Effectiveness of EfficientNet-B4 and EfficientNet-B7 Deepfake Detection Techniques

Jasmin P. Bharadiya

Ph.D. Information Technology – Cybersecurity

University of the Cumberlands

Acknowledgments

This work would not have been possible without the support of Dr. Oludotun Oni. I am especially indebted to Dr. Oludotun Oni, Committee chair of the deepfake detection dissertation, who has supported my dissertation goals and actively provided academic guidance on writing this dissertation. As my professor and mentor, he has taught me more than I could ever give him credit for here. His patience and constant effort are two key things that made this study possible for me. I am grateful to all of those I have had the pleasure to work with during this study. I would like to thank Dr. Segun Odion, a member of my committee. He has shown me direction in writing and researching machine learning. Also, very grateful to Dr. Jordan Shaw for serving as a member of my committee. His inputs on dissertation content to improve the overall merit of the study.

Abstract

The emergence of advanced deepfake technology presents a unique threat due to deepfake content's ability to evade detection and its potential to create an alternative reality. Deepfakes are realistic audio, or video renderings of an individual or group created using artificial intelligence that can match an individual's expressions and tones and portray actions or events that never really occurred, such as a presidential candidate saying defamatory remarks conducting incriminating activities. Research has indicated that deepfake technologies have rapidly advanced in efficiency and quality in the past three years. Although deepfake detection technologies are quickly becoming more sophisticated, these technologies' development appears primarily reactive and relies on sometimes outdated deepfake creation technology. This study focuses on two state-of-the-art deepfake detection techniques, EfficientNet-B4 and EfficientNet-B7, divided into two phases. In phase one, EfficientNet-B4 and EfficientNet-B7 were not trained on the latest deepfake datasets and used benchmark validation to observe their effectiveness. In phase two, training EfficientNet-B4 and EfficientNet-B7 on the latest deepfake datasets was done and validated model effectiveness with the benchmark dataset. This study aims to measure the deepfake detection efficiency of EfficientNet-B4 and EfficientNet-B7 by collecting true positives/true negatives and false positives/false negative data points from the validation test. McNemar's Chi-square test was used to evaluate detection techniques' effectiveness and found that both models had the same performance results when trained on datasets or untrained on datasets. Thus, there is no significant difference between the effectiveness of both deepfake detection techniques.

Keywords: Cybersecurity, Deepfakes, Neural Networks, National Security.

Table of Contents

Acknowledgments **2**

Abstract **3**

List of Tables **6**

List of Figures **7**

Chapter One: Introduction **8**

Overview *8*

Background and Problem Statement *9*

Purpose of Study *12*

Significance of Study *12*

Research Questions *15*

Theoretical Framework *16*

Limitations of Study *19*

Assumptions *21*

Definitions *21*

Summary *22*

Chapter Two: Review of Literature **24**

Introduction *24*

Title searches *24*

Historical Overview *25*

Current Findings *25*

Review of the Literature *27*

EfficientNet-B4 and EfficientNet-B7 ..... *46*

Identification of the Gap *50*

Summary *52*

Chapter Three: Procedures and Methodology **53**

Introduction *53*

Research Paradigm *54*

Research Design *54*

Sampling Procedure and Data Collection Sources *58*

Statistical Tests *60*

Summary *61*

Chapter Four: Research Findings **62**

Introduction *62*

Participants and Research Setting *63*

Analyses of Research Questions *65*

Summary *69*

Chapter Five: Summary, Discussion, and Implications **70**

Introduction *70*

Practical Assessment of Research Questions *70*

Limitations of Study *73*

Implications for Future Study *74*

Summary *75*

References **76**

Appendix A: IRB Approval Letter ..... Error! Bookmark not defined.

List of Tables

Table 1. 2 X 2 Accuracy of Detection before Training…… 71

Table 2. 2 X 2 Contingency Table of EfficientNet-B4 and EfficientNet-B7 before Training 71

Table 3. 2 X 2 Accuracy of Detection after Training …… 73

Table 4. 2 X 2 Contingency Table of EfficientNet-B4 and EfficientNet-B7 after Training 73

List of Figures

Figure 1.	Basic Architecture of EfficientNets ……	52
Figure 2.	The architecture of EfficientNet-B4 …….	53
Figure 3.	The architecture of EfficientNet-B7 ……..	54
Figure 4.	Phase One of the Research ……	62
Figure 5.	Phase Two of the Research: Training Efficientnet-B4 and Efficientnet-B7	63
Figure 6.	Phase Two of the Research: Validating Efficientnet-B4 and Efficientnet-B7	63
Figure 7.	McNemar's Chi-Square …….	66
Figure 8.	Phase One of the Research ……	69
Figure 9.	Phase Two of the Research: Training Efficientnet-B4 and Efficientnet-B7	70
Figure 10.	Phase Two of the Research: Validating Efficientnet-B4 and Efficientnet-B7	70
Figure 11.	McNemar's Test for Computing Chi-Square and p-value: RQ1 Results ……	72
Figure 12.	McNemar's Test for Computing Chi-Square and p-value: RQ2 Results ……	74

.

Chapter One: Introduction

Overview

Since 2010, there is evident progress in the technological world. The Internet has become a part of our lives. Everything around us requires the Internet to function. In this modern world, our social aspects have been transitioned to digital interactions more than ever before (Anderson & Rainie, 2020). This digital interaction makes us vulnerable to what we see, hear or say over the digital platform to be twisted and used against us. In that sense, it affects decision-making as we no longer can judge the authenticity of content in a sea of digital media like images, videos, or audio. Our true sense of seeing, hearing, and speaking has become impaired by a few short, potentially fake videos, and we will never know.

In 2019, approximately 4.4 billion people used the Internet, resulting in an astounding 4.4 quintillion bytes of data daily. This data includes about 95 million photos shared on Instagram daily and 400 hours of video content uploaded to YouTube every minute. A technology that utilizes the vast numbers of photos and videos uploaded to the Internet every day to create realistically, but fake audio and video renderings of an individual could be used to target political leaders or other public figures, such as executives of prominent corporations or celebrities. The diffusion of such content would result in a society that can no longer believe what it hears or sees, the precise threat represented by the emerging technology known as deepfakes.

Chapter 1 of the study focuses on the background of the deepfake to understand the threats around deepfake creation and their potential to harm society. As deepfake-generating algorithms and techniques are becoming smarter every day, developing deepfake detection techniques integrated with media platforms to safeguard them against falsified information or

content on the Internet has become crucial. Chapter one highlights the need to determine the effectiveness of deepfake detection techniques and their scope in the future.

Background and Problem Statement

Before the digital era, nations conducted denial and deception campaigns to create a false reality that deceived individuals into choosing a desired political or military outcome. United States Army sent erroneous signal transmissions to the German Army during World War II. It also established a fake unit comprised of inflatable tanks, military vehicles, plywood structures, and various decoys. The United States Army strategically deployed these units to deceive the German Army and cause them to believe they encountered a substantial United States force. The distraction provided by the fake unit allowed the troops to advance into strategic positions with limited enemy resistance (Hammond, 1994).

If inflatable tanks had the potential to influence military conflicts, the impact of an effective deception campaign using perfected deepfake technology on democracy could be immense. People's distrust of what they hear or see could create an alternative reality as people will likely choose the information that most closely aligns with their personal opinions. This alternative reality could divide society by appealing to people's biases and removing a common understanding of the truth.

Individuals' online presence since 2010 has increased as social media applications and multimedia entertainment options have multiplied. These applications rely on monetization based on the activity of online audiences and their interests and needs. Because this information has become a crucial asset for social media platforms, collecting this data relies on vague consent forms buried in service agreements that users habitually click through. Consent involves capturing our faces and bodies, a condition that demands greater transparency. On top of that,

deepfake-generating tools use online data to create deepfake content, influencing the targeted and non-targeted audience's decision-making.

Deepfakes are realistic audio or video renderings of a person or group portraying actions that never really occurred and created using artificial intelligence (AI). These renderings are becoming nearly indistinguishable from actual video or audio. Deepfakes go beyond creating photoshopped images and face-swapping applications to create videos that match the depicted person’s expressions and tone. For example, in 2018, comedian Jordan Peel appeared in a public service announcement video on the threat of deepfakes that defined former President Barack Obama (Silverman, 2018). The realistic deepfake video showed former President Obama saying out-of-character phrases that Peel initially spoke. Although this video had several noticeable flaws, it demonstrated the potential threats posed by more advanced deepfake technology.

One can detect synthesized videos can by exploiting the face-warping artifacts that result from the deepfake production pipeline. For an efficient running time, current deepfake algorithms create composite facial images of fixed sizes. These faces undergo an affine transformation (i.e., scaling, rotation, and shearing) to match the poses they replace. However, Li stated, the facial region and surrounding regions in the original image/video frame present artifacts or a resolution inconsistency due to such transformations after the compression step to generate the final image or video frames (Li & Lyu, 2019). This study, therefore, proposes to determine the effectiveness of detection techniques based on a convolutional neural network (CNN) model to detect such artifacts from the detected face regions and their surrounding areas (Li & Lyu, 2019).

Chapter one also explores how deepfake technology has impacted people's decision-making ability due to the prevalence of deepfake media content on the Internet. The study

observes deepfake's capabilities in terms of manipulating content like audio, video, or images through artifacts such as lip-syncing, head puppetry, eye blinking, and so on. It is essential to determine the effectiveness of deepfake detection techniques based on specific artifacts and technology to compare the detection techniques.

Deepfake detection leverage multi-modal detection techniques to determine whether target media has been manipulated or synthetically generated. Existing detection techniques consist of manual and algorithmic methods. Manual techniques include human media forensic practitioners who use software tools. Algorithmic detection uses an AI-based algorithm to identify manipulated media (Jaiman, 2020).

As society has moved into the digital era, the capabilities of denial and deception campaigns have advanced. These campaigns have evolved past inflatable tanks and plywood structures to curate coordinated online movements to spread false, deceptive, or malicious content. Russia has received significant attention over the past decade to use cyber technology to conduct disinformation campaigns undermining Western national security. The Russian government has used Internet trolls employed by its Internet Research Agency (IRA) to spread disinformation through social media posts. Multiple intelligence agencies have accused the country of meddling in the 2016 United States presidential election and the 2017 French presidential election, among other political objectives (United States Senate Committee on Foreign Relations, 2018).

In a 2017 translated interview between an alleged former Russian IRA employee and a Russian news network, the employee stated, "We did not have the goal of turning the Americans towards Russia. Our goal was to turn the Americans against their government. Cause unrest, displeasure, downgrade Obama's rating" (Kotlyar, 2017, p. 1). Combatting deepfake technology

has become necessary, driving governments and tech giants to develop ways to mitigate deepfake threats using AI. Artificial intelligence has had promising outcomes, but AI versus AI can be challenging because both the generation and detection of deepfakes are forms of AI. As deepfake-generating algorithms and techniques become increasingly innovative, detection techniques rely on data trained on older versions of deepfakes.

Though how effective it is to use AI for deepfake detection is valid, there are currently few alternative solutions for slowing the rise of deepfake content. Moreover, as the tools make deepfakes advance, they can distinguish them from authentic content wanes (HPE, 2019). Falsified videos created by AI algorithms on CNNs, and deep neural networks (DNNs) are a contemporary twist to the alarming online disinformation problem. However, digital pictures and videos' creation and influence are not new (Farid, 2012).

Purpose of the Study

The purpose of this study is to perform a comparative analysis of the different AI deepfake detection techniques to determine their effectiveness. In addition, this study determines if there is a significant difference between the EfficientNet-B4 and EfficientNet-B7 deepfake detection techniques when trained on the latest deepfake datasets and when they are not. Both research questions focus on evaluating machine learning models' performance using McNemar's Chi-square test. This study will help determine the current and future capabilities of deepfake technologies, their scope, and the threat deepfake poses to national security.

Significance of the Study

Deepfakes initially used to swap celebrity faces with individuals in pornographic videos, deepfake technology has significantly advanced in recent years. Researchers have developed highly sophisticated artificial intelligence programs that have resulted from an increasing degree

of accuracy when creating deepfakes (Dack, 2019). However, there are still simple applications of this technology, such as downloadable face-swapping programs. These deepfake technologies pose an actual potential national security threat driven by advanced algorithms and unsupervised learning.

The creation of deepfakes relies on artificial intelligence: neural networks and generative adversarial networks (GANs). A neural network learns similarly to the human brain. The more photos and videos of a person exposed to a neural network, the more accurately it can replicate their mannerisms, tones, and facial expressions when creating a deepfake (Dack, 2019). The previously discussed prevalence of social media in today's digital society provides a constant and growing learning material source for a neural network.

Deepfake detection approaches with significant development observed in the past two years. DNNs in the latest times have made the progression to make credible false videos progressively simpler and quicker. DNN made fraudulent videos initially caught the public's attention in 2017 when a Reddit account named “deepfakes” began posting artificial pornographic tapes produced using a DNN-based face-swapping algorithm. Afterward, the term “deepfake” more broadly refers to any AI-produced mimicking videos. It involves a slide of real detection techniques advanced in less than two years, mainly created on deep learning (Afchar et al.,2018; Li et al., 2018); the accessibility of numerous extensive deepfake video datasets (Guera et al., 2018; Matern et al., 2019; Yang et al., 2019); and two public challenges keen to deepfake detection, specifically, the DARPA MFC18 Synthetic Data Detection Challenge and the Facebook deepfake Detection Challenge.

Generative methods that produce novel samples from high-dimensional data distributions, such as images, are finding widespread use (Karras et al., 2018). For example, speech synthesis,

image-to-image translation, and image imprinting (Karras et al., 2018). Currently, the most prominent approaches are autoregressive models, variation autoencoders (VAE), and generative adversarial networks (GAN) (Karras et al., 2018).

Numerous present approaches are prepared as frame-level binary classification problems; these techniques fall into three main types. The first type created on irregularities showed in the physical/physiological features in the deepfake videos.

- Head puppetry involves creating a video of a marked person's full head and upper shoulder using a source person's video, so the completed target looks similar to the source.
- Face swapping contains producing a video of the target with the expressions replaced by the source's created faces though keeping the identical facial expressions.
- Lip syncing forms a fabricated video by only influencing the lip area so that the target speaks somewhat that s/he does not say.

Currently, they all have significant strengths and weaknesses (Karras et al., 2018). Autoregressive models –such as PixelCNN –produce sharp images but are slow to evaluate and do not have a latent representation (Karras et al., 2018). Furthermore, they directly model the conditional distribution over pixels, potentially limiting their applicability (Karras et al., 2018). Further into the literature by Karras et al. (2018), GANs and the training are stable in large resolutions; there is a long way to true photorealism. However, semantic sensibility and understanding dataset-dependent constraints, such as particular objects being straight rather than curved, leaves a lot to be desired. There is also room for improvement in the microstructure of the images (Karras et al., 2018). Existing deepfake detection approaches mainly focus on face-

swapping videos, which report for the leads of deepfake videos spread on a virtual network (Karras et al., 2018).

The continued advancements in face-swapping techniques will result in more challenging deepfake videos, which the existing algorithms will be harder to detect (Korshunov & Sébastien, 2018). Korshunov also stated new databases and new, more generic methods developed in the future. In addition, subjective evaluations to study the vulnerability of human subjects to deepfakes are also required. Possibly, a new arms race between deepfake methods and detection algorithms has begun (Korshunov & Sébastien, 2018).

The development of deepfake detection techniques based on how deepfakes are generated and trends promote creating deepfakes in the first place. To measure the benefits and limitations of different deepfake detection techniques like Visual Geometry Group (VGG), DNN, CNN and GAN would help future studies develop other detection techniques. It may also help guide legislation to protect our online data and users, potentially our democratic rights. In recent years DARPA and big internet companies are making joint efforts to develop, research, and publish better AI deepfake detection techniques. Provides open-source training data for new models to continue developing AI technologies.

Research Questions

Deepfake detection techniques are developed based on generative processes. Target-specific audio/video content features can be categorized based upon a compound detection technique, such as image processing, speech synthesis, and so on. Although there has been progressing in detecting deepfake videos, several concerns remain over the compound detection method that requires caution (Lyu, 2020).

RQ 1: Is there a difference in the effectiveness of the state-of-the-art CNN deepfake detection techniques of EfficientNet-B4 and EfficientNet-B7 when untrained on any datasets and validated on the same benchmark dataset?

H_{01}: There is no significant difference between the effectiveness of EfficientNet-B4 and EfficientNet-B7 techniques when untrained on any datasets and validated on a benchmark dataset.

H_{a1}: There is a significant difference between the effectiveness of the EfficientNet-B4 and EfficientNet-B7 techniques when untrained on any datasets and validated on a benchmark dataset.

RQ 2: Is there a difference between the effectiveness of the state-of-the-art CNN deepfake detection techniques of EfficientNet-B4 and EfficientNet-B7 when trained on similar datasets and validated on the same benchmark dataset?

H_{02}: There is no significant difference between the effectiveness of the EfficientNet-B4 and EfficientNet-B7 techniques when trained on similar datasets and validated on a benchmark dataset.

H_{a2}: There is a significant difference between the effectiveness of the EfficientNet-B4 and EfficientNet-B7 techniques when trained on similar datasets and validated on a benchmark dataset.

Theoretical Framework

A video's temporal consistency is not imposed efficiently and enforced in the detection process to use the video's spatiotemporal content to detect deepfakes (Nguyen et al., 2020). Integrating conventional networks and recurrent units manipulate the frames' temporal inconsistencies (Sabir et al., 2019). Recurrent units use long short-term memory (LSTM) and

CNN to extract temporal features of the video, denoted through the sequence of frames. The deepfake detection network takes the sequence frames and estimates the frame sequence's possibilities, either original or deepfakes. The CNN extracts the frame features and then feeds them into the LSTM to generate temporal sequence frames (Delp & Güera, n.d.; Echizen; Nguyen et al., 2020; Yamagishi, 2018). The CNN is developed at a fixed resource budget and then scaled up for better accuracy if more resources are available (Tan & Le, 2019a).

The conventional practice for model scaling is to arbitrarily increase the CNN's depth or width or use larger input image resolution for training and evaluation. While these methods improve accuracy, they usually require tedious manual tuning but often yield a suboptimal performance. A novel model scaling method uses a simple yet highly effective compound coefficient to scale CNN's more structured manner. Unlike conventional approaches that arbitrarily scale network dimensions such as width, depth, and resolution, this method uniformly scales each dimension with a fixed set of scaling coefficients. As a result, these newly developed models, called EfficientNets, surpass state-of-the-art accuracy with up to 10 times better efficiency (Tan & Le, 2019b).

EfficientNets is relatively new, and it is essential to determine their effectiveness in terms of deepfake detection as they are now widely used to develop deepfake detection techniques. Therefore, this study focuses on 1) two exemplary EfficientNet deepfake detection techniques to determine their effectiveness and accuracy with the latest datasets, 2) trends of deepfake creation and 3) a trajectory for the future development deepfake detection techniques based on EfficientNet.

The quality of deepfakes has been improving, and detection algorithms' performance must be enhanced in kind. Beyond advances in AI that make deepfakes hard to detect, current

detection techniques typically emphasize the weaknesses of deepfake creation tools. Still, this approach cannot always discover the deepfakes created by these tools (Wubet, 2020). As such, deepfake detection techniques should be trained on old and new deepfake training datasets simultaneously for the best results. Models validated against benchmark dataset to understand the scope of such EfficientNet based detection techniques and their effectiveness on actual data.

This study uses contingency theory of decision making to measure the efficiency of EfficientNet-B4 and EfficientNet-B7 state-of-the-art deepfake detection techniques. The effectiveness of which machine learning model is best depending upon various aspects of the situation, such as the amount of relevant information during the machine learning testing, the importance of decision quality, and acceptance as to how much true positives can make a difference and validate the accuracy of the model. In this study, to measure two state-of-the-art deepfake detection models: EfficientNet-B4 and EffiicientNet-B7 effectiveness by comparing the performance of machine learning models in two research settings. One is comparing EfficeintNet-B4 pre-trained vs. pre-trained EfficientNet-B7. Second is EfficeintNet-B4 & EfficientNet-B7 models trained on deepfake datasets then tested on deepfake benchmark dataset. The researcher believes the Contingency theory of decision making is best suitable for making a strategic decision on which deepfake detection model outperformed in both scenarios. Results from both research questions will give a better understanding of machine learning models as count of true positives, false positives, true negatives & false negatives can provide accuracy of both EfficientNet-B4 & EffcientNet-B7 models in terms of deepfake detection. Results will help evaluate which machine learning model is best for deepfake detection based on their performance on the deepfake benchmark dataset, detection accuracy, training time, machine learning models' convolutional neural layer performance, etc.

Limitations of the Study

The lack of literature on specific deepfake detection techniques might affect evaluating these techniques and their potential. Individual researchers build detection techniques for a feature of deepfake content, such as spatial clues, temporal clues, generalizable forgery detection, spoofing forgery detection, target-specific face forgery, and other areas. However, this specificity creates a high dependency on the quality and quantity of training datasets to accurately detect feature-based deepfakes in the real world. In addition, certain external factors can limit the detection of deepfake videos, and several concerns over this detection method require caution (Lyu, 2020).

Quality of deepfake Datasets

The provision of largescale datasets of deepfake videos enables deeper consideration of deepfake detection methods. However, Lyu's (2020) study of deepfake videos in existing datasets reveals stark contrasts in visual quality compared to deepfake videos circulated online.

Performance Evaluation

Currently, detecting deepfake videos is often formulated, solved, and evaluated as a binary classification problem in which each video is categorized as real or deepfake. This dichotomy is helpful for controlled experiments to develop and test deepfake detection algorithms using videos that are either pristine or made with deepfake generation algorithms (Lyu, 2020). However, the situation is more complicated when any detection method is deployed on the Internet. For example, videos may be fabricated or manipulated in other ways beyond deepfakes; thus, not being detected as a deepfake video does not necessarily indicate that the video is real (Lyu, 2020). Additionally, a deepfake video may be subject to other manipulations, and one label might not comprehensively reflect the possible manipulations (Lyu, 2020).

Furthermore, in a video with multiple subjects' faces, just one or a few faces are generated with deepfake for a fraction of the frames (Lyu, 2020). In such a case, the binary classification scheme must be extended to multiclass, multi-label, and native classification/detection to handle the complexities of deepfake (Lyu, 2020).

Justification of Detection Results

Current deepfake detection methods perform batch analysis over an extensive collection of videos. Detection results of a video generated by a synthesis algorithm are not as valuable for practitioners if they do not conform to similar trend deepfakes used in training. It is typical to justify the analysis's numerical score for publication or use in court in such scenarios. One can only understand the authenticity of detection results if training datasets and benchmark datasets have similar generative deepfake methods; that is, a model trained on fruit images cannot detect vegetables accurately. However, many data-driven deepfake detection methods lack information, especially those supporting DNNs (Lyu, 2020).

Temporal Aggregation

Most existing deepfake detection methods support a binary classification at the frame level by determining whether a private frame is real or fake. Although simple and easy, there are two problems with this technique.

Firstly, one must ignore temporal consistency among frames as (i) many deepfake videos exhibit temporal artifacts and (ii) real or deepfake frames tend to occur at continuous intervals (Lyu, 2020). Secondly, it necessitates an additional step when a video-level integrity score is needed, and the scores of individual frames need to be aggregated to compute such a score (Lyu, 2020).

Social Media Laundering

To avoid wasting network bandwidth and safeguard users' privacy, these videos are usually stripped of metadata, downsized, and then heavily compressed before being uploaded to social media platforms (Lyu, 2020). Unfortunately, these operations, commonly called social media laundering, harm the ability to recover traces of underlying manipulation and, at the same time, increase false positive detections, or the classification of an actual video as a deepfake(Lyu, 2020).

Assumptions

Scholarly research on deepfakes is sparse. This study focuses on filling the literature gap by undertaking a programming experiment to measure the effectiveness of two state-of-the-art deepfake detection techniques, EfficientNet-B4 and EfficientNet-B7, by training detection models on similar datasets and validating models benchmark testing dataset. The proposed experimental environment and training/testing datasets will produce a reliable accuracy ratio.

Definitions

In addition to variables and constructs, this study used several uncommon terms. This segment provides short definitions or descriptions of these terms.

Artificial intelligence: machines programmed to think like humans and mimic their actions (*beyond the camera news*, 2018).

Autoregressive models: models that predict future behavior based on past behavior (Smith, 2015).

Convolutional neural network (CNN): deep learning algorithm which can take an input image, assign importance (learnable weights and biases) to various aspects/objects in the picture, and differentiate one from the other (Saha, 2018).

DARPA MFC18 Synthetic Data Detection Challenge: for image and video forensics technologies supported by the United States government (Delgado, 2017).

Deep neural network (DNN): deep learning is a sub-field of machine learning in AI that deals with algorithms inspired by the brain's biological structure and functioning to aid machines with intelligence (Moolayil, 2019).

deepfakes: fake videos or audio recordings that look and sound authentic (Porup, 2019).

Facebook deepfake Detection Challenge: an open, collaborative initiative to accelerate the development of new technologies for detecting deepfakes and manipulated media (Pesenti, 2019).

Generative adversarial networks approach generative modeling using deep learning methods, such as convolutional neural networks (Brownlee, 2019).

Speech synthesis: artificial simulation of human speech with a computer or other device (Techopedia, 2019).

Unsupervised learning is a machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses (Mathworks, 2020).

Summary

Chapter 1 presented the background of deepfakes and their impact, summarizing the many types of deepfake detection techniques, particularly target-specific deepfake detection that focuses on a single problem. Deepfake is a relatively new technology that can be an active part of everyday life through entertainment and education (Griffin, 2019). However, Deepfake also affects decision-making, which poses numerous threats to society. Therefore, it is necessary to review the available literature on deepfakes to gather more insight. Chapter 2 reviews current

research on deepfake detection techniques and their development to understand the scope of deepfake creation and detection techniques. Chapter 3 focuses on methodology, research approach, and design. Chapter 4 presents comparative study findings of detection techniques and their effectiveness based on unseen training datasets to calculate an accurate detection ratio. The final chapter interprets and discusses the study results.

Chapter Two: Review of Literature

Introduction

New technologies are being developed and becoming a massive part of humankind's day-to-day lifestyle, where digital technologies make it increasingly difficult to distinguish between real and fake media. One of the most recent developments contributing to the problem is the emergence of deepfakes (Westerlund, 2019). These hyper-realistic videos apply AI to depict someone say and do things that never happened (Westerlund, 2019).

In recent years, fake news has become an issue that threatens public discourse, human society, and democracy (Westerlund, 2019). Fake news refers to fictitious news-style content fabricated to deceive the public (Aldwairi & Alwahedi, 2018; Jang & Kim, 2018). False information spreads quickly through social media, impacting millions of users (Figueira & Oliveira, 2017). Presently, one out of five Internet users gets their news via YouTube, second only to Facebook (Anderson, 2018).

This rise in video popularity highlights the need for tools to confirm media and news content authenticity, as novel technologies allow effective video manipulation (Westerlund, 2019). Furthermore, given the ease in obtaining and spreading misinformation through social media platforms, it is increasingly hard to know what to trust, which results in harmful consequences for informed decision-making, among other things (Westerlund, 2019).

Title searches

This study utilized current 2019-2020 peer-reviewed articles via various search engines like Google and safari. These search results pointed to scholarly, including Google AI Blog, IEEE, ELSEVIER, Facebook AI Lab, etc. Also utilized Google Scholar to locate open access articles. Used the following search terms to find articles specific to this study: deepfake

detection, AI deepfake latest findings, neural networks, deepfake datasets, neural network model for image classification, and so on. Variations of these terms are used to ensure detailed search results.

Historical Overview

Indeed, today we live in what some have called a "post-truth" era characterized by digital disinformation and information warfare led by malevolent actors running false information campaigns to manipulate public opinion (Westerlund, 2019). It's called "deepfakes," hyper-realistic videos using face swaps that leave little trace of manipulation (Wagner & Henderson, 2020). Deepfakes are the product of artificial intelligence (AI) applications that merge, combine, replace, and superimpose images and video clips to create fake videos that appear authentic (Wagner & Henderson, 2020).

The game-changing factor of deepfakes is the scope, scale, and sophistication of the technology involved, as almost anyone with a computer can fabricate fake videos that are practically indistinguishable from authentic media (Westerlund, 2019).

Current Findings

In this study, deepfakes are categorized based upon their detection feature and approach to facilitate further research objectives of the comparative study of different deepfake detection techniques. Moreover, to learn about detection technique's scope and measure effectiveness of other deepfakes detection techniques as an accurate ratio of detecting feature in deepfakes testing dataset. In recent literature where, Generative adversarial networks (GANs) (Goodfellow et al., 2014) have enabled a set of different face manipulations techniques. Including identity (Karras et al., 2019), (Majumdar et al., 2019), facial attributes (Wang et al., 2018), as well as facial expressions (Jiang et al., 2020; Liu et al., 2019). While several manipulation detection

approaches are image-based (Afchar et al., 2018; Rössler et al., 2019), others are targeted towards videos (Agarwal et al., 2019; Li et al., 2018; Sabir et al., 2019) or jointly on audio videos (Korshunov et al., 2018).

As noted in the literature that although some video-based approaches might perform better than image-based ones, such methods are only applicable to kinds of attacks. For example, many of them (Agarwal et al., 2019; Li et al., 2018;) may fail if the eye area's quality is sufficiently good or the synchronization between video and audio is natural. Image-based approaches are general-purpose detectors; for instance, they apply to steganalysis and facial reenactment video detection to detect computer-generated images, which later extended to detecting computer-manipulated images (Wang & Dantcheva, 2020). However, the performance of such approaches on new tasks is limited compared with that of task-specific algorithms (Khodabakhsh et al., 2018; Rössler et al., 2019). This comparative study will show that such algorithms are challenged. Suppose confronted with videos outside of the training data.

To justify the hypotheses of this study, theoretical or/and conceptual framework is as follows:

- State-of-the-art deepfake detection CNN methods, which have claimed high accuracy ratio in deepfake detection feature of action recognition, will be selected.
- After shortlisting CNN-based deepfake detection techniques, all work to detect the same feature as action recognition or face swap later be categorized by detection feature.
- They use open-source deepfake datasets, such as Celeb-DF (Li et al., 2019), to train deepfake detection models and later use different testing datasets such as fractions (Li et al., 2019). The unseen testing dataset for a model. The later model will give accurate detection of deepfake and real videos found in the testing dataset.

- The current study will be performed on different deepfake detection models sharing the same detection feature. It will include training all detection models with the same training dataset and measuring their accuracy through the same testing unseen dataset.

Review of the Literature

Existing deepfake detection approaches mainly focus on face-swapping videos, which report for the leads of deepfake videos spread on the virtual network. Numerous present methods as frame-level binary classification problems; these techniques fall into three main types. Techniques in the first type created on irregularities showed in the physical or physiological features in the deepfake videos. The work technique misuses the inspection that many deepfake videos lack reasonable eye blinking due to online pictures as training data, which typically do not have closed eyes for aesthetic purposes. Illogical head postures in deepfake videos are a clue to uncover deepfake videos. The characteristic interactive patterns of an individual. The interval sequences of facial markers from real videos as input to detect deepfake videos.

The second type of deepfake detection algorithm uses signal-level objects presented throughout the creation method. Also, as produced faces get merged into the original video frames, DNN splicing detection techniques can help. The third type of deepfake detection technique is data-driven, which straightforwardly engages various neural networks trained on real and deepfake videos but catching authentic artifacts. Finally, the lip-sync-based detection system aims to distinguish real video, where lip movement and speech are synchronized, from tampered video, where lip movements and audio, which may not necessarily be speech, are not synchronized (Korshunov & Sébastien, 2018).

In most deepfake detection techniques, long short-term memory (LSTM) network is used and trained to separate tampered and non-tampered videos (Chung et al., 2016). In this system, a

technique used is MFCCs as audio features (Le & Odobez, 2016) and distances between mouth landmarks as visual features (Suwajanakorn et al., 2017). Principal Component Analysis (PCA) is applied to the joint audio-visual features to reduce the dimensionality of the blocks of features (Korshunov & Sébastien, 2018). First, the lip-syncing-based algorithm cannot detect face swapping, as GANs can generate facial expressions with high quality that can match audio speech.

Therefore, currently, only image-based approaches are capable of detecting deepfake videos effectively. Second, the combination of Image Quality Metrics (IQM) + Support Vector Machine (SVM) system has a high accuracy of detecting deepfakes. However, videos generated with the High Quality (HQ) model pose a more severe challenge (Korshunov & Sébastien, 2018). It means that a more advanced technique for face swapping will be even more challenging to detect (Korshunov & Sébastien, 2018). Third, the proposed convolutional neural network with a Y-shaped auto-encoder demonstrated its effectiveness for classification and segmentation tasks without using a sliding window, commonly used by classifiers. Information sharing among the classification, segmentation, and reconstruction tasks improved the network's overall performance, especially for the mismatch condition for seen attacks (Nguyen et al., 2019).

Moreover, the auto-encoder can quickly adapt to unseen attacks using only a few samples for fine-tuning. The gap in this is the effect of using residual images on the auto-encoders performance, processing high-resolution images without resizing, improving its ability to deal with unseen attacks, and extending it to the audio-visual domain by Nguyen et al. (2019).

GANs compute feature statistics from individual images and across the mini-batch, thus encouraging the batches of generated images and training images to show similar statistics (Karras et al., 2018). Later adding a mini-batch layer towards the end of the discriminator layer

learns a large tensor that projects the input activation to various statistics. Finally, each concatenated the layer's output so that the discriminator can use the statistics internally (Karras et al., 2018).

Advances in digital human faces have become the basis for modern facial image editing tools (Rössler et al., 2019). Moreover, it can split the editing tools into two main categories: identity modification and expression modification; aside from manually editing the face using Photoshop tools, many automatic approaches proposed in the last few years.

To detect synthesized videos by exploiting the face warping artifacts resulted from the deepfake production pipeline. First, the current deepfake algorithms create composite face images of fixed sizes for an efficient running time. Then, these faces underwent an affine transform (i.e., scaling, rotation, and shearing) to match the poses of the target faces that they will replace (Li & Lyu, 2019). As such, the facial region and surrounding regions in the original image/video frame will present artifacts, the resolution inconsistency due to such transforms after the subsequent compression step to generate the final image or video frames (Li & Lyu, 2019). Therefore, this study proposes determining the effectiveness of detection techniques based on a Convolutional Neural Network (CNN) model to detect artifacts from the detected face regions and their surrounding areas (Li & Lyu, 2019).

Further into this chapter is how deepfake has impacted people's decision-making ability exposed to the deepfake media content flooded on the internet in various situations. By observing deepfake unfolding its capabilities in terms of what can manipulate way authenticity of the content. To narrower down to what media content features like audio, video, or image are being tampered with what type of artifacts such as lip-syncing, head puppetry, eye blinking, etc. It is essential to determine the effectiveness of deepfake detection techniques based on specific

artifacts and technology used in the first place to draw a comparison between two deepfake detection techniques.

Deepfake detection includes solutions that leverage multi-modal detection techniques to determine whether target media has been manipulated or synthetically generated. Existing detection techniques manual and algorithmic methods. Manual techniques include human media forensic practitioners, often armed with software tools. Algorithmic detection uses an AI-based algorithm to identify manipulated media. (Jaiman, 2020).

In the detection process, the video's temporal consistency is not imposed efficiently and enforced to use the video's Spatiotemporal contents to detect deepfakes (Nguyen et al., 2020). Integrating conventional networks and recurrent units manipulate the frames' temporal inconsistencies (Sabir et al., 2019). Using both Long Short-Term Memory (LSTM) and CNN can extract temporal features of the video denoted through the sequence of frames. The deepfake detection network takes the sequence frames and estimates the sequence's possibilities, either original or deepfake. CNN is used to extract the frame features and then fed them into the LSTM to generate temporal sequence frames (Delp & Güera, n.d.; Echizen; Nguyen et al., 2020; Yamagishi, 2018).

CNN developed at a fixed resource budget and then scaled up for better accuracy if more resources are available (Tan & Le, 2019a). The conventional practice for model scaling is to arbitrarily increase the CNN depth or width or use larger input image resolution for training and evaluation. While these methods improve accuracy, they usually require tedious manual tuning and still often yield suboptimal performance. A novel model scaling method uses a simple yet highly effective compound coefficient to scale CNN's more structured manner. Unlike conventional approaches that arbitrarily scale network dimensions, such as width, depth, and

resolution. A method that uniformly scales each dimension with a fixed set of scaling coefficients. This newly developed family of models called EfficientNets surpass state-of-the-art accuracy with up to 10x better efficiency (Tan & Le, 2019b).

As efficiencies are relatively new and determine their effectiveness in terms of deepfake detection is necessary, as it is now widely used in the development of deepfake detection by researchers. Therefore, this study focuses on two states of the art EfficientNet deepfake detection techniques to determine their effectiveness and accuracy with the latest datasets, trends of deepfake creation, and a learning curve for the future development of deepfake detection techniques based on EfficientNet.

The deepfakes' quality has been improving, and detection algorithms' performance requires to be enhanced consequently. Developing better detection methods to combat fakes generated by deepfakes is challenging because artificial intelligence, which powers deepfakes, is rapidly growing, making deepfakes hard to detect. Currently, detection techniques typically emphasize the weaknesses of the deepfake creation tools. This approach is not always functional to detect deepfakes created by all deepfake creation tools (Wubet, 2020). So, for best results, deepfake detection techniques are trained on old and new deepfake training datasets simultaneously and later validated with benchmark validation testing dataset as moving forward to understand the scope of such EfficientNet based detection model their effectiveness against actual data. As if it can implement active development of detection techniques on media platforms as soon as possible.

The most common identity editing technique is face swapping, for instance, face swap filters in Snapchat, which has gained significant popularity as lightweight systems can now run on mobile phones (Rössler et al., 2019). Now, facial reenactment techniques are available, which

alter a person's expressions by transferring the facial expressions of a source person to the target (Rössler et al., 2019). Although impressive progress in detecting deepfake videos, there l concerns over these detection methods that Lyu (2020) cautioned.

Quality of deepfake Datasets.

The provision of largescale datasets of deepfake videos is an enabling consider the event of deepfake detection method. However, a better study of the deepfake videos in existing datasets reveals some stark contrasts in visual quality to the deepfake videos circulated on the net (Lyu, 2020).

Performance Evaluation.

Currently, detecting deepfake videos is often formulated, solved, and evaluated as a binary classification problem, where each video is categorized as real or deepfake. However, such dichotomy is simple to line up in controlled experiments, test deepfake detection algorithms using videos that are either pristine or made with deepfake generation algorithms.

However, the image is murkier when the detection method is deployed on the planet. For example, videos may be fabricated or manipulated in ways aside from deepfakes, so not being detected as a deepfake video does not necessarily suggest the video footage may be a real one. A deepfake video is subject to various manipulations, and one label might not comprehensively reflect such (Lyu, 2020).

Furthermore, during a video with multiple subjects' faces, just one or some are generated with deepfake for a fraction of the frames that the binary classification scheme must be extended to multiclass, multi-label, and native classification/detection to handle the complexities of planet media imitations (Lyu, 2020).

Justification of Detection Results

A synthesis algorithm's detection results are not as valuable if they are not validated with similar deepfakes used in training. For practitioners, it is typical to use the analysis's numerical score to justify it being accepted for publishing or employed in court. The authenticity of detection results can only be ensured if training datasets and benchmark datasets are developed through similar generative deepfake methods as those used in the videos: models trained on fruit images cannot detect vegetables accurately. However, many data-driven deepfake detection methods that support deep neural networks usually lack explanation due to the appending nature of the DNN models (Lyu, 2020).

Temporal Aggregation

Most existing deepfake detection methods are supported binary classifications at the frame level, determining the likelihood of an individual frame being real or a deepfake. Although simple and easy, there are two problems with this technique.

First, the temporal consistency among frames can ignore, as (i) many deepfake videos exhibit temporal artifacts and (ii) real or deepfake frames tend to exist in continuous intervals. Second, it necessitates an additional step in which a video-level integrity score is needed: the scores over individual frames must be aggregated to compute such a score (Lyu, 2020).

Social Media Laundering

Videos are usually stripped of meta-data, downsized, and then heavily compressed before uploading to social platforms to avoid wasting network bandwidth and safeguard users' privacy. Unfortunately, these operations, commonly called social media laundering, are detrimental to recovering traces of underlying manipulation and increasing false positive detections, that is, classifying an actual video as a deepfake (Lyu, 2020).

So far, deepfake detection methods that use amplitude features suffer from social media laundering (Lyu, 2020). A practical measure to enhance the robustness of deepfake detection methods to social media laundering is to dynamically integrate simulations in training data and improve datasets to include video performance, both real and synthesized (Lyu, 2020).

The continued advancements in face-swapping techniques will make deepfake harder to detect by the existing models (Korshunov & Sébastien, 2018). In addition, subjective evaluations to study the vulnerability of human subjects to deepfakes are also required. Thus, a new arms race between deepfake methods and detection algorithms may have begun (Korshunov & Sébastien, 2018).

Deepfakes result from generative adversarial networks (GANs), two artificial neural networks that work together to create real-looking media (Westerlund, 2019). These two networks, called “the generator” and “the discriminator,” are trained on the same dataset of images, videos, or sounds. Scientists are also discovering the use of GANs to detect anomalies in X-rays and create virtual chemical compounds to tweak materials science and medical discoveries (Westerlund, 2019).

Businesses are interested in brand-applicable deepfake technology's potential because it can transform e-commerce and advertising significantly. For example, brands can contract with supermodels, create their deepfakes and show fashion outfits on various fashion models with different skin colors, heights, and weights. Deepfakes even allow for super personal content that turns consumers into models (Westerlund, 2019). In addition, the technology enables virtual fitting, allowing consumers to preview how an outfit would look on them before purchasing it and generate targeted fashion ads that vary depending on the time, weather, and viewer (Westerlund, 2019). An obvious potential use is trying on clothes online quickly; the technology

allows people to create digital clones and have these personal avatars travel across e-stores. It also enables people to digitally try on bridal gowns or suits and then virtually experience a wedding venue (Westerlund, 2019). Furthermore, AI can provide unique artificial voices that differentiate companies and products to make branding distinction easier.

There are few major categories of deepfake creators: deepfake hobbyists, politicians such as foreign governments, other malicious actors such as fraudsters (Shanoob, 2020), and real actors, such as television companies. Unfortunately, it is difficult to track down individuals in deepfake hobby communities. For example, after one user introduced celebrity porn deepfakes to Reddit in late 2017, a newly founded deepfake hobbyist community only took a few months to reach 90,000 members (Wagner & Henderson, 2020).

Many hobbyists focus on porn-related deepfakes, while others place famous actors in films they never acted in to produce comedic effects. Their deepfakes intent is to be entertaining, funny, or politically satirical and help them gain social media followers (Shanoob, 2020). Some hobbyists may be acting out of more concrete personal motivations, such as raising awareness about the potential of deepfake technology or getting deepfake-related paid work, for example, with music videos or television shows (Westerlund, 2019).

While meme-like deepfakes by hobbyists can amuse social media users, certain political players with agendas, including political activists, radicals, and foreign nations, can use deepfakes in propaganda campaigns to manipulate public opinion and undermine confidence in a country's institutions (Westerlund, 2019). For warfare, deepfakes are weaponized disinformation aimed at interfering with elections and sowing civil in warfare unrest. In addition, we can anticipate more and more domestic and foreign-state-funded internet "troll farms" that use AI to create and deliver fake political videos tailored to social media users' specific biases.

Deepfakes are also increasingly deployed by fraudsters to conduct stock market manipulation and commit various other financial crimes (Westerlund, 2019). For example, criminals have already used AI-generated fake audios to impersonate an executive and ask for an urgent cash transfer. The visual materials required to produce impersonations of executives are often available on the internet. Deepfake technology can use executives' visual and audio impersonations from, for example, TED Talk videos available on YouTube.

Deepfakes pose a significant threat to our society, political systems, and businesses because they require journalists to filter real from fake news. In addition, they threaten national security by disseminating propaganda and interference in elections. They also hamper citizen trust toward information issued by authorities, and they raise cybersecurity issues for people and organizations (Westerlund, 2019).

The journalism industry will likely have to face a massive consumer trust issue due to deepfakes. Deepfakes pose a more significant threat than "traditional" fake news because they are harder to spot, and people are inclined to believe the fakes are real (Simon, 2019). In addition, the technology allows the production of seemingly legitimate news videos that place journalists' reputations and the media at risk (Westerlund, 2019). To access video footage shot by a bystander of an incident can provide a competitive lead to a newscast channel.

Misattributed video footage such as an actual protest march or suggests captioned to suggest it happened somewhere else is a growing problem. The rise of deepfakes will augment this issue. For example, while looking for eyewitness videos about the mass shooting in Christchurch, New Zealand, Reuters came across a video that claimed to show the moment a suspect was shot dead by police (Westerlund, 2019). However, they quickly learned it was from a different incident in the USA and not the suspect in the Christchurch shooting.

U.S. intelligence bureaucrats have repetitively alerted about the threat of foreign interfering in American politics, specifically in the lead-up to elections. The intelligence community is worried that deepfakes will be used to threaten national safety by distributing political propaganda and disrupting election campaigns (Westerlund, 2019). Putting words in somebody's mouth on an audiovisual that goes viral is an effective weapon in today's fake news wars, as such altered videos can easily skew voter opinion (Westerlund, 2019).

Another threat of deepfakes is how continuous interaction with misinformation affects people (Westerlund, 2019). Deepfakes are likely to hamper digital knowledge and citizens' trust in authority-provided information, as fake videos showing government officials saying things that they never did make people doubt authorities (Westerlund, 2019). The feeling that much information, including video, cannot be trusted has resulted in a phenomenon called "information apocalypse" or "reality apathy." People may even dismiss actual footage as fake simply because they have become embedded in the notion that whatever they do not want to consider must be fake (Westerlund, 2019). The threat is not an instance of the deception of people but that they will come to regard everything as a deception.

Cybersecurity issues form additional threats imposed by deepfakes. As a result, the corporate world expressed interest in protecting itself against viral frauds (Westerlund, 2019). These can include deepfakes used for market and stock manipulations that may show a company official speaking racist or misogynistic slurs, stating a fake merger, making false announcements about financial losses or bankruptcy, or committing a crime (Westerlund, 2019).

Deepfaked porn or product broadcasts for brand disruption, blackmail, or humiliate management (Westerlund, 2019). In addition, deepfake technology enables the real-time digital impersonation of an executive, for example, asking an employee to perform a vital cash transfer

or provide confidential data. Further, deepfake technology can generate a fake identity (Westerlund, 2019). In live-stream videos, the ability to convert an adult face into a child's or younger person's face has raised concerns about using the technology by child predators (Westerlund, 2019). Lastly, deepfakes can contribute to the increase of malicious scripts. Researchers found that a site dedicated to deepfakes misused their users' computers to mine cryptocurrencies (Westerlund, 2019).

Most deepfakes on social platforms like YouTube or Facebook are harmless fun or artistic works that use dead or living public figures (Westerlund, 2019). However, there are also examples from the dark side of deepfakes, namely celebrity and revenge porn and attempts at political and non-political influencing (Westerlund, 2019). Most such deepfakes are goofs, pranks, and funny memes with comedic or satirical effects. Deepfakes like Nicolas Cage acting in movies he has never starred in, such as Indiana Jones or Terminator 2 (Westerlund, 2019).

Some enjoyable instances of deepfakes include a video that swaps Alden Ehrenreich to young Harrison Ford in clips from Solo: A Star Wars movie and a video of actor Bill Hader's appearance on Late Night David Letterman (Westerlund, 2019). While Hader speaks about Tom Cruise, his face changes into Cruise's. Some deepfakes show dead celebs, such as the band Queen's ex-vocalist Freddie Mercury's face imposed on actor Rami Malek and the Russian mystic Grigori Rasputin singing Beyoncé's powerful ballad "Halo" (Westerlund, 2019). An art museum in the US has used the technology to bring Salvador Dali back to life to greet visitors. Alternative AI system makes anybody dance like a prima ballerina by imposing areal dancer's steps onto a target person's body, thereby creating a video that displays the target as a professional dancer (Westerlund, 2019).

Examples of harmful deepfakes, however, are increasingly popping up. Deepfake technology enables celebrity and revenge porn—involuntary pornography using celebrities' and laypeoples' images shared on social networks. Scarlett Johansson has appeared in deepfaked adult movies in which their faces are superimposed over porn stars' faces. In 2018, a deepfake generated by Hollywood filmmaker Jordan Peele performed as previous U.S. President Obama deliberated fake news threats and mocking present president Trump (Westerlund, 2019).

In 2018, Donald Trump's deepfake video offered Belgium advice about climate change (Westerlund, 2019). The video was created by the Belgian political party "sp. a" to draw attention from people to sign up for an online petition calling on the Belgian government to take more vital climate action. The video triggered outrage about the American president meddling in a foreign country like Belgium's climate policy (Westerlund, 2019). The U.S. Democratic Party deepfaked its chairman Tom Perez in 2019 to highlight the potential threat of deepfakes to the 2020 election (Westerlund, 2019).

While these are examples of limited political influence, other deepfakes may have a more lasting impact. For instance, in Central Africa in 2018, a deepfake of Gabon's long-unseen president Ali Bongo, who was believed to be in poor health or dead, was cited as the trigger for an unsuccessful coup by the Gabonese military (Westerlund, 2019). Likewise, a viral clip deepfake of a man confessing to having sex with a local cabinet minister in Malaysia caused political controversy (Westerlund, 2019).

Non-political individuals have also been the targets of deepfakes. A high-quality deepfake was released in June 2019 by two British artists featuring Facebook CEO Mark Zuckerberg, who racked up millions of views (Westerlund, 2019). The video misleadingly represents Zuckerberg giving respect to Spectre, a fictional evil group from the James Bond

sequences that teaches him how to take total control of billions of peoples' confidential data and thus own their futures (Westerlund, 2019). Because it contains news footage, deepfake technology, and a voice actor, the video was meant to show how one can use technology to manipulate data (Westerlund, 2019).

At present, civil and criminal law ignore deepfakes. However, legal experts have suggested adapting current rules to cover libel, defamation, identity fraud, or impersonating a government official using deepfakes.

Virginia state law of revenge porn recently made sharing "falsely created" images and videos a misdemeanor and prolonged the rule to include deepfakes (Westerlund, 2019). In addition, people's visual depictions in deepfake videos are not exact copies of any existing material but rather new representations generated by AI. Thus, regulators must navigate a complicated legal landscape around free speech and ownership laws to regulate deepfake technology use properly.

A proper legal resolution to the proliferation of destructive deepfakes would not be a complete ban on the technology (Westerlund, 2019), as that would be unethical. Today's social media companies enjoy liberty for users' content on their platforms (Westerlund, 2019). One legislative way could be to walk back social media companies' legal distance from their users' content, thus making users and platforms more responsible for posted material (Westerlund, 2019). However, legislation has had negligible influence on malicious actors such as foreign states and terrorists, who may run massive misinformation campaigns in contradiction of other states' social media platforms (Westerlund, 2019).

Company policies and voluntary action may offer more efficient tools against deepfakes (Westerlund, 2019). For example, politicians can commit to not using illicit digital campaign

tactics or spreading disinformation such as deepfakes in their election campaigns ("Singapore Introduces New Bill to Punish Online Fake News Outlets," 2021). In addition, social media companies need to enforce ethics even though disruptive content pushed to the top of the feed is money-wise a success because it expands advertisement engagement time (Westerlund, 2019). Unfortunately, few social media firms have policies about deepfakes.

Instead, they down-rank it to make it more difficult to find less prominent in users' news feeds (Westerlund, 2019). On the other hand, the increases in hate speech, fake news, and disinformation polluting digital platforms have led some firms to take more action, such as suspending user accounts and investing in faster detection technology (Westerlund, 2019).

Facebook cuts off any content recognized as fabricated or deceptive by third-party fact-checkers from running ads and making money. The corporation fact-checks parties, academics, specialists, and officials to find novel resolutions (Westerlund, 2019). For example, Instagram's algorithms will not recommend people view content marked as "false" by fact-checkers (Westerlund, 2019). The Wall Street Journal and Reuters have formed corporate teams to train their reporters to identify fake content and adopt detection techniques like cross-referencing locations on Google maps and reverse image searching among news media companies. Education and training are vital for fighting deepfakes. Despite significant news coverage and fears presented by authorities, the threat of deep fakes has not yet been reckoned with by the public (Westerlund, 2019).

In general, there is an essential need to increase public mindfulness about AI's potential for exploitation (Westerlund, 2019). As deepfakes provide cybercriminals with new social engineering tools, companies and organizations must alert and establish cyber resilience plans.

Critical thinking and digital learning are not taught in institutes. Yet, these traits contribute to children's abilities to spot fake news and interact more respectfully with each other online.

Younger people should promote these skills among the older, less technology-savvy population. The intention for this is the public needs to critically evaluate the genuineness and social perspective of a video and the trustworthiness of its source (Westerlund, 2019). Thus, to understand the video's actual intent. It is also crucial to consider that quality is not an indicator of a video's genuineness (Westerlund, 2019).

The best technique for saving photos and videos is to have strategies to safeguard them. Many government bodies & companies use facial recognition technology and storing vast amounts of digital data for security purposes (Westerlund, 2019). Generally, the challenges of using technology to validate the content and recognize fakes relate to doing it at scale. There are far more resources and people working on developing technology to create deepfakes than technology to detect them.

Twitter scraps 8 million accounts a week to spread content through manipulative strategies (Westerlund, 2019). It is a massive challenge for technologies to go through all posted material in a short time. Further, deepfake developers tend to use published deepfake research results to improve their technology and get around new detection systems. For instance, scientists found that initial deepfake approaches failed to mimic the frequency at which a person was blinking (Westerlund, 2019). Recent programs fixed the lack of blinking or unnatural blinking after the findings were published.

The Defense Advanced Research Projects Agency (DARPA) has significant corporate opportunities for private cybersecurity companies to develop solutions for deepfake detection, build reliable policies, weed out unlawful bots, and fight against fraud and digital pollution

(Westerlund, 2019). Yet, the improvement of anti-deepfake technology alone is not enough. Organizations must also adopt these technologies (Westerlund, 2019). They can modernize the government in any given country, giving it the ability to face and help protect its citizens against deepfakes ("Singapore Introduces New Bill to Punish Online Fake News Outlets," 2021).

While it is more difficult for people to distinguish between real videos and fakes, AI can detect deepfakes (Westerlund, 2019). Media forensic experts have suggested subtle indicators that help detect deepfakes. These include a range of deficiencies, such as face vary, flicker, and alteration; waviness in a person's actions; discrepancies with talking and mouth movements; irregular movements of fixed objects such as mic stands; contradictions in lighting, reflections, and shadows; blurry boundaries; angles and blurring of facial features; an inconsistency in breathing; unnatural eye direction; missing facial characteristics like the known mole on a cheek; smoothness and heaviness of clothing and hair; too smooth skin; lacking hair and teeth particulars; misalignment in face proportion; irregularities in pixel levels; and strange behavior like an individual doing something implausible (Westerlund, 2019).

Artificial intelligence algorithms can analyze photo response non-uniformity (PRNU) patterns in the footage, defects unique to the light sensor of particular camera models, or biometric blood flow data indicated by progressive changes on a person's face in a video (Westerlund, 2019). In addition, new fake detection algorithms are based on mammalian auditory systems, such as how mice detect inconsistencies and subtle audio mistakes that are often ignored by humans (Westerlund, 2019).

Artificial Intelligence can determine if a video has been manipulated by examining videos on a frame-by-frame basis to track signs of fake or assessment an entire video at once to inspect soft biometric critical signatures, including irregularities in head movements, speech

patterns, ad facial expressions (Suciu, 2020). It can tailor the latter method for individuals, such as high-profile politicians, who are likely to be deepfaked.

The problem with deepfakes is proving that something is false and proving that an object is authentic. Video authentication is essential to news media companies, who must spread videos in a trustless environment. Details of the video's creator and circulation may be hard to trace (Westerlund, 2019). A simple practice is authenticating content with digital watermarks as digital forensic techniques (Westerlund, 2019). It would be ideal for creating a "truth layer," an automated system across the internet that would measure fake versus authentic. To have a way like an authentication process, every video posted to a social media site would go through (Westerlund, 2019).

For example, smartphone software can digitally identify a film's footage (Westerlund, 2019). Upon footage playback, this watermark matched with the unique fingerprint to check for a match and specify the viewer with a score that indicates the likelihood of tampering (Westerlund, 2019). Digital watermarking, like hashing, can provide an audiovisual file with a short string of lost digits if the video is tampered with (Westerlund, 2019). It can also offer authenticated alibis for public figures, given that they constantly record where they are and what they are doing.

Video legitimacy is also provided by mapping a video's provenance, whether the footage originates from a reputable source and how it has traveled online since (Westerlund, 2019). Blockchain technology can help verify the origins and sharing of videos by producing and storing the digital key signature in a ledger that is nearly impossible to manipulate (Westerlund, 2019). Nonetheless, people will always choose not to believe a verification tool and still desire to consume and endorse fake media.

Finally, technology can prevent the making of deepfakes by injecting "noise" into photos or videos. This noise is hidden to the human eye but prevents the visual material from being used in automated deepfake software (Westerlund, 2019). One could also wear specially designed 3D-printed specs to avoid facial recognition by misleading deepfake software into misclassifying the wearer. This technology could help probable targets such as politicians, celebs, and executives prevent deepfakes from being made (Westerlund, 2019).

In addition, researchers developing GAN technologies can design and put proper safeguards in place so that their technologies become more challenging to misuse for disinformation purposes. Like the cybersecurity domain in general, the first step towards a solution is understanding the problem and its ability to affect us – only then does it become possible to develop and apply technical tricks that can solve the challenges (Westerlund, 2019). That said, no technological solutions can entirely remove the risk of deepfakes. Every problem has a technical solution that may take the discussion away from more existential questions of why deepfakes exist and other threats AI can impose (Suciu, 2020). The most efficient ways to fight against deepfakes from spreading are a mixture of legal, educational, and sociotechnical advances.

Technical solutions for deepfake detection, content authentication, and deepfake prevention form a dynamic field of security methods (Westerlund, 2019). Subsequently, there are numerous opportunities for technology entrepreneurs, especially in cybersecurity and AI. In addition, national defense agencies and big IT companies acknowledge the ongoing efforts in developing deepfake detection techniques (Suciu, 2020). These include small-scale and large-scale deepfake detection techniques and approaches and brainstorming ideas to benefit national security and democracy in general.

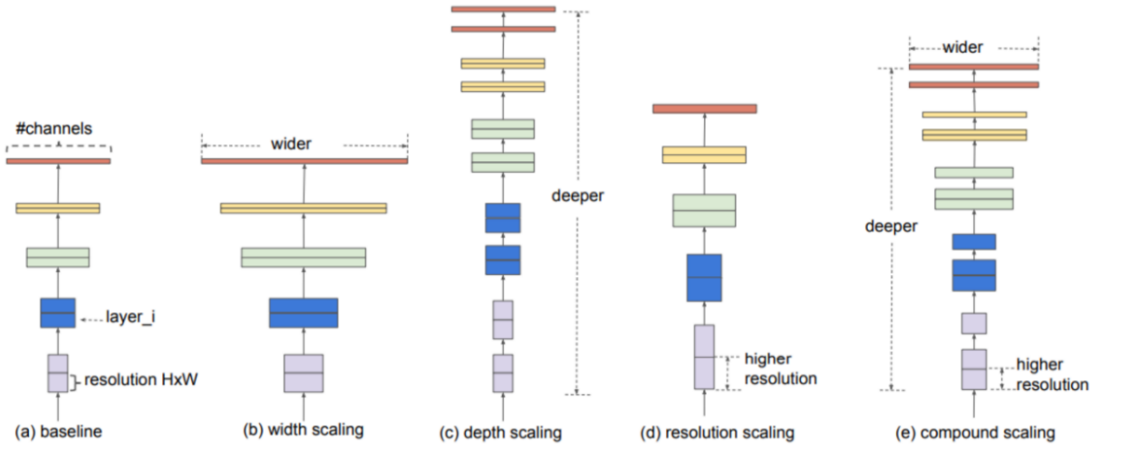
This comparative study will contribute to the body of knowledge on deepfakes by filling the gap in the literature. Currently, there is minimal discussion regarding the evaluation of new deepfake detection techniques or holding a standard for the most acceptable outcome of deepfake detection techniques. The bar should be based on specific detection features and their effectiveness, which can also benefit the future efforts of deepfake detection approaches.

EfficientNet-B4 and EfficientNet-B7

EfficientNets, introduced by Google AI, were proposed as more efficient neural network techniques to improve state-of-the-art results (Tan & Le, 2019). Generally, the neural network models are too broad, deep, or with a very high resolution. Increasing these characteristics initially helps the model, but it quickly saturates because it has made more parameters and becomes inefficient. In EfficientNet, parameters are scaled more principled; gradually, everything is increased (Agarwal, 2020), as shown in Figure 1 below.

Figure 1

Basic Architecture of EfficientNets)



Note. Comparison of different scaling methods (Tan & Le, 2019, p. 2)

EfficientNet models, from EfficientNet-B0 to EfficientNet-B7, are based on the compound scaling method (Tan & Le, 2019), which uses the compound coefficient ϕ to uniformly scale network width, depth, and resolution in a principled way:

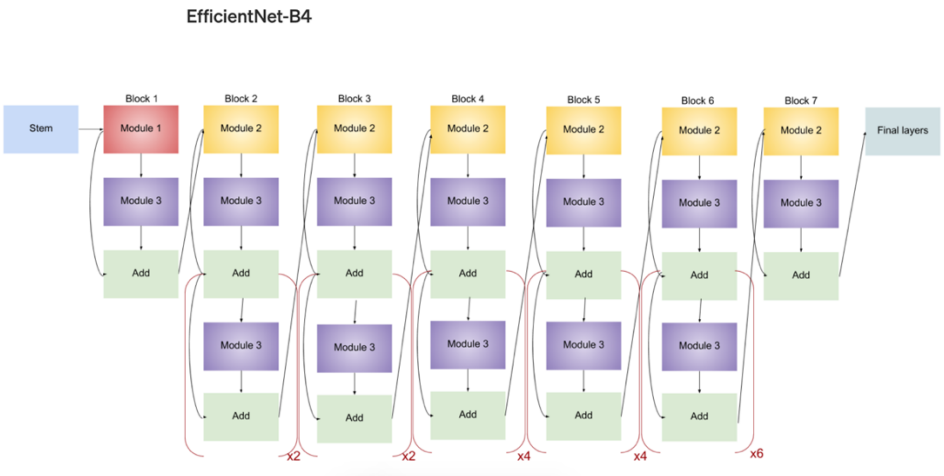
$$depth\ d = \alpha^{\phi}, width\ w = \beta^{\phi}, resolution\ r = \gamma^{\phi}$$

$$s.t.\ \ \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2, where\ \alpha \geq 1, \beta \geq 1, \gamma \geq 1$$ (Tan & Le, 2019).

In which *α, β,* and *γ* are constants that a small grid search can determine. Intuitively, *φ* is a user-specified coefficient that controls how many more resources are available for model scaling, while *α, β,* and *γ* specify how to assign these extra resources to network width, depth, and resolution, respectively (Tan & Le, 2019). Figure 2 shows the base architecture of state-of-the-art EfficientNet-B4 regarding compound scaling in the neural layers of EfficientNets.

Figure 2

Architecture of EfficientNet-B4

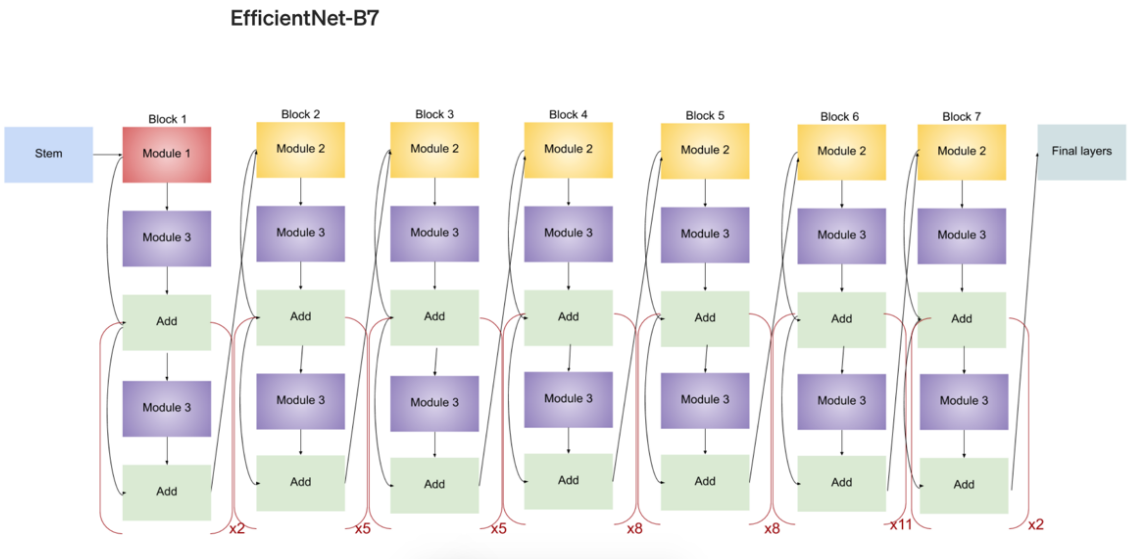


Note. The basic architecture of EfficientNet-B4 scaling (Agarwal, 2020)

In general, all scaling methods improve accuracy with the cost of more FLOPS (Tan & Le, 2019). The compound scaling method can improve accuracy by up to 2.5% over other single dimension scaling methods of convolutional neural networks (Tan & Le, 2019). The architecture of EfficientNet-B7 consists of more neural network layers, as shown in Figure 3.

Figure 3

The architecture of EfficientNet-B7



Note. The basic architecture of EfficientNet-B7 scaling (Agarwal, 2020)

EfficientNet-B7 architecture has more depth, width, and resolution than the previous EfficientNet model series. This study follows the theory of slack resources and their impacts by analyzing the performance of both state-of-the-art EfficientNet-B4 and EfficientNet-B7 models and their accuracy on the latest datasets. As machine learning models depend on the datasets they have been trained on, training datasets become the first knowledge ground of any machine learning model and determine the model's ability to predict anomalies in the future.

Deepfake generation improves by the day, making models underperform until they are trained on new trends and improve their ability to detect various features simultaneously,

creating a race between deepfake generation and deepfake detection. Deepfake datasets can be sensitive; they might not be available as a resource for future enhancement of detection techniques. Delays in the availability of datasets and the lack of understanding of deepfake trends can impact such entities' decision-making processes.

Methodology and instrumentation/data sources/research materials

Currently, new levels of the manipulation of images and videos have surfaced. To generate a video with these new manipulation levels, one must superimpose a target person's face onto a source person's video. This method is widely accepted and referred to as a deepfake. Most recent research on deepfakes has created forged videos based on a short video of the source person and a single ID photo. Also, fully synthesized audio video images could replicate a target person's synchronous speech and lip movements.

The model is the more challenging training setting for CNN-based deepfake detection (as indicated by lower related detection rates). High-quality compressed data was generally detected at rates ranging between 70.97% and 95.73% (XceptionNet). Intuitively low-quality compressed data had the lowest detection rates, at 55.98% to 81% (XceptionNet). The methodologies used in prior research (Wang & Dantcheva, 2020) included:

- single manipulation techniques, in which all algorithms' performances were trained and tested on single manipulation techniques (reports accurate classification rates [TCR])
- cross-manipulation designs, which is the most challenging setting because it is unlikely to provide knowledge on whether and how videos are manipulated to explain whether and how exploited videos will be provided (reports TCR); and

- all manipulation techniques, which are CNNs in comparison to image-forgery detection algorithms (reports TCR).

Convolutional neural networks are developed with a fixed resource budget and scaled up for better accuracy if more resources are available (Tan & Le, 2019a). The conventional practice for model scaling is to arbitrarily increase the CNN depth or width or use larger input image resolutions for training and evaluation. While these methods improve accuracy, they usually require tedious manual tuning and still often yield suboptimal performance. A novel model scaling method uses a simple yet highly effective compound coefficient to scale CNN's more structured manner. Unlike conventional approaches that arbitrarily scale network dimensions, such as width, depth, and resolution, this method uniformly scales each dimension with a fixed set of scaling coefficients. As a result, this newly developed family of models, EfficientNets, surpasses state-of-the-art accuracy with up to 10 times better efficiency (Tan & Le, 2019b).

EfficientNets are relatively new and determining their effectiveness in terms of deepfake detection. They are now widely used in the development of deepfake detection by researchers. However, they are necessary. This study focuses on two state-of-the-art EfficientNet deepfake detection techniques to determine their effectiveness and accuracy with the latest datasets and trends of deepfake creation and establish a learning curve for the future development of deepfake detection techniques EfficientNets.

Identification of the Gap

The automated generation and manipulation of audio, images, and video provide inspiring perspectives for science, art, video production, video animation, special effects, and reliving already passed actors. While highly intriguing from a computer vision perspective, deepfakes entail several challenges and threats, given that they are fabricated manipulations.

In particular, decision-makers cannot trust what is seen or heard on the video, as the two predominantly trusted human innate senses, i) sight and ii) sound, are betrayed by deepfakes (Wang & Dantcheva, 2020). Our society relies heavily on producing and exchanging legitimate and trustworthy documents, so sound and images have recorded our history. They have informed and shaped our perceptions of reality (Wang & Dantcheva, 2020), and "a picture is worth a thousand words" (Martin, 2019). Social media has pushed online videos as a mainstream source of information; deepfakes threaten distorting reality. Deepfake techniques have become accessible to the public via phone applications such as FaceApp-1 and ZAO-2.

Currently, manipulations include subtle imperfections that humans cannot, however, detect. If trained well, computer vision algorithms called deepfake detection techniques can be used. These techniques use various technologies or/and various technological approaches like artificial intelligence, machine learning, deep learning, neural networks, etc. Researchers, academic resources, computer science programmers/developers, and AI enthusiasts can all come together to solve the problem of deepfake detection. This is critical today as the technologies for making deepfakes are progressively available, and social media platforms can quickly spread fake content. Sometimes, deepfakes do not need to be seen by a massive audience to cause detrimental effects.

People who create deepfakes with malicious purposes need to deliver them to their aim audiences as part of their interruption tactic without using social media. Intelligence services can utilize this approach to influence decisions made by influential people such as politicians, leading to national and international security threats. Seeing the alarming deepfake problem, the research community has focused on developing deepfake detection algorithms, and numerous results have been reported (Nguyen et al., 2020).

Summary

Chapter 2 covered a literature review on deepfake detection techniques and their development and provided detailed information on deepfakes. This chapter builds a bridge between prior efforts in the field of deepfakes, deepfake generation, detection techniques, and their benefits, and the threat deepfakes pose to society, along with future needs in deepfake research. Next, Chapter 3 will dive into the methodology, research approach, and design of this study. It will describe procedures to evaluate the effectiveness of different deepfake detection techniques based upon detection features. Chapter 4 will support comparative study findings in detection techniques and their effectiveness based upon unseen training datasets for measuring accurate detection ratios. Finally, Chapter 5 of the dissertation will have the interpretation and discussion of the study result.

Chapter Three: Procedures and Methodology

Introduction

Deepfakes are realistic audio or video renderings of a person or group created using artificial intelligence that portrays actions that never really occurred. These renderings are becoming nearly indistinguishable from the actual person or persons. Deepfakes eclipse the creation of Photoshopped images and face-swapping applications to create videos that match the depicted person's expressions and tone.

For example, in 2018, comedian Jordan Peele helped produce a public service announcement on the threat of deepfakes, which depicted former President Barack Obama overlaid on a comedy video (Silverman, 2018). The seemingly realistic deepfake video presented the former president saying out-of-character phrases spoken by the comedian. Although this video did have several noticeable flaws, it demonstrated the potential future threats posed by more advanced deepfake technology.

As deepfake generating algorithms and techniques become more innovative daily, detection techniques rely on trained data from older deepfake versions. Today, however, there are not many viable solutions for slowing the rise of deepfake content. In this regard, a fundamental question arises of how effective it is to use AI for deepfake detection.

Falsified videos created by AI algorithms, CNNs are a contemporary twist to the alarming online disinformation problem. However, the creation of digital pictures and videos—and the implementation of their influence—are not new (Farid, 2012). As tools, advance deepfake, our ability to distinguish them from authentic content will wane (HPE, 2019), with the human outsmarted by the decision-making abilities of intelligent technology.

Research Paradigm

A quantitative study is appropriate for evaluating deepfake detection techniques. In this study, each deepfake detection model is trained on a similar training dataset and tested on the same testing dataset to reveal how effective the detection techniques are compared to one another. The same features of deepfake (such as eye blinking, head puppetry, lip-syncing, or face-swapping) can be measured in percentages.

To determine the categories of deepfake detection techniques and understand the latest trends in creating deepfake, publicly available deepfake training datasets are used, such as FaceForensics ++ (Rössler et al., 2019). Additionally, the EfficientNet-B4 and EfficientNet-B7 deepfake detection models used in this study are publicly available (Seferbekov, 2020) to encourage researchers to optimize such techniques' accuracy or gaps.

This quantitative study has aimed to evaluate the effectiveness of different deepfake detection techniques by categorizing them into distinct detection features that methods claim can be successfully detected. Deepfake detection development prioritizes big-tech companies and defense agencies because deepfake has become a threat to national security and democracy. Therefore, findings can drive efforts to create, develop, and brainstorm detection techniques; validation of detection techniques' rapid development is becoming more crucial to improve further and increase the deepfake detection accuracy ratio.

Research Design

Deepfake detection development prioritizes big-tech companies and defense agencies, as deepfake has become a threat to national security and democracy. Therefore, it drives efforts to create, develop, and brainstorm detection techniques; it becomes more crucial to validate the rapid development of detection techniques to improve and increase the deepfake detection

accuracy ratio. This quantitative study has aimed to evaluate the effectiveness of different deepfake detection techniques by categorizing a specific class of detection aspects that methods claim can be detected.

This study's approach for validating deepfake detection techniques, using EfficientNet-B4 and EfficientNet-B7, exists in the two phases of research questions one and two. The focus of validating both deepfake detection models on unseen benchmark datasets without training them; this research approach resembles an actual environment setup. For example, suppose deepfake detection models are part of one of the digital platforms where training on the latest trend is impossible because machine learning models detect deepfakes on the go. In that case, real-time models will not quickly learn from new trends in deepfake, as no training has occurred. In this case, detection models EfficientNet-B4 and EfficientNet-B7 are expected to know and detect the latest deepfake trends. This phase's result is crucial, as it has provided the prediction of performance without training based on the latest deepfake datasets potentially unavailable for training.

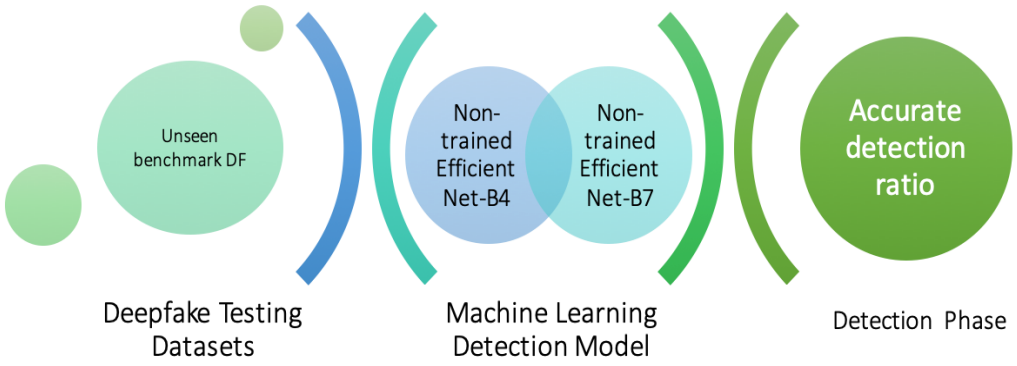
Also, validating both deepfake detection models on unseen benchmark datasets after training them on the latest deepfake datasets; this research approach resembles the supervised learning of machine learning models. If deepfake detection models are part of a digital platform where training on the latest trend is possible, then the EfficientNet-B4 and EfficientNet-B7 detection models are expected to learn and detect the latest deepfake trends. This phase's result is crucial because it predicts performance with training. The latest deepfake datasets are present for training, and the latest deepfake generation trends are accessible for further development of models.

Research Procedure for RQ 1

To answer RQ 1, the effectiveness of state-of-the-art CNN deepfake detection techniques EfficientNet-B4 and EfficientNet-B7 when untrained on any datasets and validated on the same benchmark testing dataset is established. This validation test offers an accurate detection ratio of EfficientNet-B4 and EfficientNet-B7 before training. Later, EfficientNet-B4 and EfficientNet-B7 on unseen benchmark deepfake datasets are conducted as presented in Figure 4.

Figure 4

Phase One of the Research



Note. The flow of the first research phase, untrained EfficientNet-B4 & EfficientNet-B7 put on unseen benchmark testing

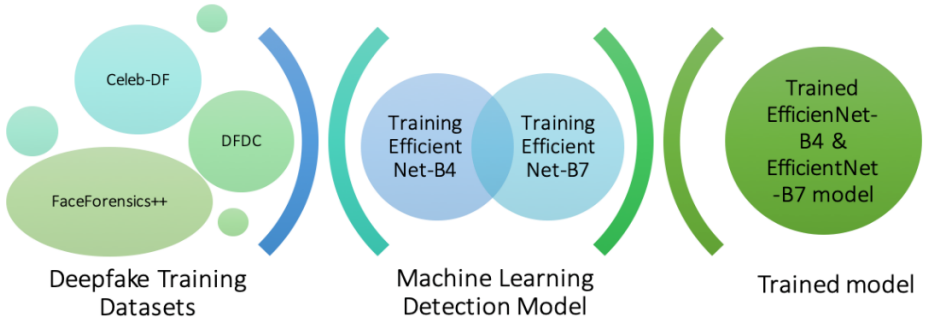
Research Procedure for RQ 2

To answer RQ 2, the effectiveness of state-of-the-art CNN deepfake detection techniques EfficientNet-B4 and EfficientNet-B7 when trained on deepfake datasets and validated on the same benchmark testing dataset is established. This validation test provides an accurate detection ratio of EfficientNet-B4 and EfficientNet-B7 after training both detection models on various deepfake datasets, such as CelebDF (Li et al., 2019), deepfake Detection Challenge (DFDC; Facebook, 2020), and FaceForensics++ (Rössler et al., 2019). Later, EfficientNet-B4 and

EfficientNet-B7 on Unseen Benchmark deepfake datasets are conducted, as revealed in Figures 5 and 6 below.

Figure 5

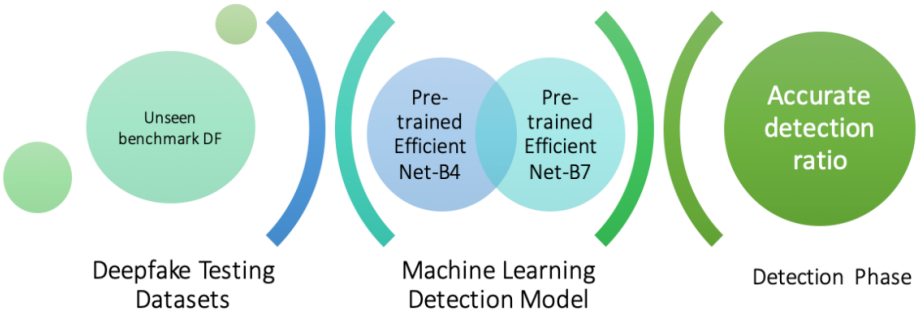
Phase Two of the Research: Training EfficientNet-B4 and EfficientNet-B7



Note. The flow of second research phase, EfficientNet-B4 & EfficientNet-B7 training on Celeb-DF, FaceForensics++ and DFDC datasets

Figure 6

Phase Two of the Research: Validating EfficientNet-B4 and EfficientNet-B7 Unseen Benchmark Dataset



Note. The flow of the second research phase, trained EfficientNet-B4 & EfficientNet-B7 put on unseen benchmark testing

The researcher measured features using the same testing data to evaluate each model's effectiveness in deepfake detection. The researcher created the unseen deepfake dataset by reserving 30% of videos from Celeb-DF (Li et al., 2019) and the DFDC (Facebook, 2020) datasets alongside FaceForensics++ (Rössler et al., 2019) benchmark dataset of 1,000 images.

The Unseen Benchmark dataset consists of 1,692 videos from Celeb-DF (Li et al., 2019), 30,000 videos from DFDC (Facebook, 2020), and 1,000 images from the FaceForensics++ benchmark dataset (Rössler et al., 2019) to equal 32,692 validation data points, which are not used as part of the training dataset for the EfficientNet-B4 and EfficientNet-B7 models. Deepfake detection techniques on the testing dataset provide an accuracy ratio of how effective these detection techniques are when recognizing new deepfakes, a practical approach for the comparative study of different deepfake detection techniques. In addition, it can further guide the development of detection approaches, with merit-based evaluation influencing the standard acceptability of deepfake detection techniques for future use.

Sampling Procedure and Data Collection Sources

This comparative study intends to benefit the development of future deepfake detection models to fight against deepfake; It has aimed to measure the effectiveness of deepfake detection techniques and thus provide a significant contribution to deepfake literature. Datasets for training deepfake detection models were collected from open sources and utilized for the validation testing dataset. A fraction of 30% from the original dataset was reserved as an unseen benchmark dataset from each deepfake data collection.

Test results were collected and analyzed through the Statistical Package for the Social Sciences (IBM SPSS, 2020) tool to perform data analysis using McNemar's Chi-Square test. In this experiment, no population was involved as datasets used in this study are secondary data. Instead, the respective sources collected data as part of a large-scale experiment organized by Facebook (Facebook, 2020). In addition, other datasets, FaceForensics++ (Rössler et al., 2019) and Celeb-DF (Li et al., 2019), were also part of this research to create a validation dataset for models trained and tested upon various detection measures. This also prevents deepfake

detection models from overfitting during training, so the accuracy-test results performed with the benchmark dataset can be effective.

Quantitative Sample Size

For this study sample, dataset one was Celeb-DF (Li et al., 2019), consisting of 5,639 deepfake videos, corresponding to more than two million frames. The real source videos are based on publicly available YouTube video clips of 59 celebrities of diverse genders, ages, and ethnic groups. The second sample dataset was the deepfake Detection Challenge dataset, consisting of a face swap video dataset. Over 100,000 total clips were sourced from 3,426 paid actors, produced with several deepfake GAN-based and non-learned methods. Finally, the third dataset was FaceForensics++ (Rössler et al., 2019), consisting of 1,000 original video sequences manipulated with four automated face manipulation methods: deepfakes, Face2Face, FaceSwap, and NeuralTextures.

For validation, the Unseen deepfake dataset was created from Celeb-DF (Li et al., 2019), DFDC (Facebook, 2020), and FaceForensics++ (Rössler et al., 2019) by reserving a set of 30% videos from Celeb-DF (Li et al., 2019), the DFDC (Facebook, 2020) dataset, and the FaceForensics++ (Rössler et al., 2019) benchmark dataset of 1,000 images. The total of the Unseen benchmark dataset equals 32,692 validation data points, with Celeb-DF's 1,692 videos (Li et al., 2019), DFDC's 30,000 videos (Facebook, 2020), and the FaceForensics++ benchmark dataset's 1,000 images (Rössler et al., 2019).

For this study, the three primary datasets were Celeb-DF (Li et al., 2019), DFDC (Facebook, 2020), and FaceForensics++ (Rössler et al., 2019) for training both EfficientNet-B4 and EfficientNet-B7 deepfake detection models. The benchmark dataset was created by reserving 30% of data from each dataset, this study utilized whole available datasets to validate machine

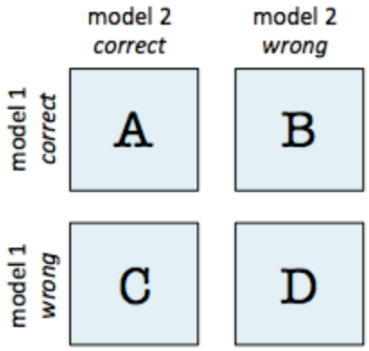
learning model performance on the similar trend of deepfake. There is no population involved, as all three datasets, Celeb-DF (Li et al., 2019), DFDC (Facebook, 2020), and FaceForensics++ (Rössler et al., 2019), were made available as open-source for research purposes. All datasets were used, as the training machine learning models required a batch of data from datasets on certain epochs to train and learn. Later, both EfficientNet-B4 and EfficientNet-B7 models were validated on a reserved 30% of the original datasets as a benchmark dataset.

Statistical Tests

For data analysis, the classification accuracy of EfficientNet-B4 and EfficientNet-B7, measured by McNemar's test, is based on a two times two contingency table of the two models' predictions as revealed in Figure 7. McNemar's test was calculated using the SPSS (IBM SPSS, 2020) tool to thoroughly evaluate each deepfake detection model's effectiveness. McNemar's test uses statistics to compare the two models' effectiveness as output data of detection models are binary data (as in deepfake detected Yes or deepfake detected No).

Figure 7

Model 1-EfficientNet-B4, and Model 2-EfficientNet-B7 in McNemar's Chi-Square



Note. The Basic structure of McNemar'schi-square contingency table

McNemar's chi-square is a suitable statistical test for machine learning models and more convenient for a classification problem like deepfake detection; thus, detection techniques are binary (e.g., an image or video content is deepfake or actual). In McNemar's contingency table,

test results consist of true positives, false positives, true negatives, and false negatives. For this study, The Researcher trained EfficientNet-B4 and EfficientNet-B7 detection models on certain epochs of 100, which means training was performed in batches of 100 data points (100 frames of video). Several hours were required to finish training both models on the same datasets. They collect validation test results through the unseen benchmark dataset encompassed at least three hours for each detection model. Results were in the format of binary: real 0 and deepfake 1.

Summary

Chapter 3 has conducted a deep dive into the methodology, research approach, and design and procedures to help evaluate the effectiveness of different deepfake detection techniques based on state-of-the-art CNNs, EfficientNet-B4 and EfficientNet-B7. Chapter 3 has presented a profound part of this study, as we developed research procedures for the effectiveness of EfficientNet-B4 and EfficientNet-B7 through two research phases. As a result, Chapter 4 supports comparative study findings in detection techniques and their effectiveness based upon an unseen training dataset to measure the accurate detection ratio. The deepfake detection ratios for EfficientNet-B4 and EfficientNet-B7 are presented in the next chapter. The validation test's findings are understood through McNemar's chi-square ratio of true positives, false positives, true negatives, and false negatives. The final chapter of this dissertation introduces the interpretation and discussion of the study results.

Chapter Four: Research Findings

Introduction

Deepfakes surpass the creation of Photoshopped images and use face-swapping applications to create videos that match the depicted person's expressions and tone. In this study, an accuracy test has been conducted for state-of-the-art deepfake detection techniques, EfficientNet-B4 and EfficientNet-B7, through a validation benchmark dataset that was created from Celeb-DF (Li et al., 2019), DFDC (Facebook, 2020), and FaceForensics++ (Rössler et al., 2019). A set of 30% videos was reserved from Celeb-DF (Li et al., 2019) and the DFDC (Facebook, 2020) dataset alongside FaceForensics++ (Rössler et al., 2019) benchmark dataset of 1,000 images. The Unseen Benchmark dataset of 1,692 videos from Celeb-DF (Li et al., 2019), 30,000 videos from DFDC (Facebook, 2020), and 1,000 images from FaceForensics++ benchmark dataset (Rössler et al., 2019) equals 32,692 validation data points. Deepfake datasets are secondary data for this study, as deepfake datasets are available online. For further research into deepfake detection techniques, the state-of-the-art EfficientNet-B4 and EfficientNet-B7 machine learning models have been adapted (Melas-Kyriazi, 2019).

Analyses of Research Questions and Hypotheses Statements

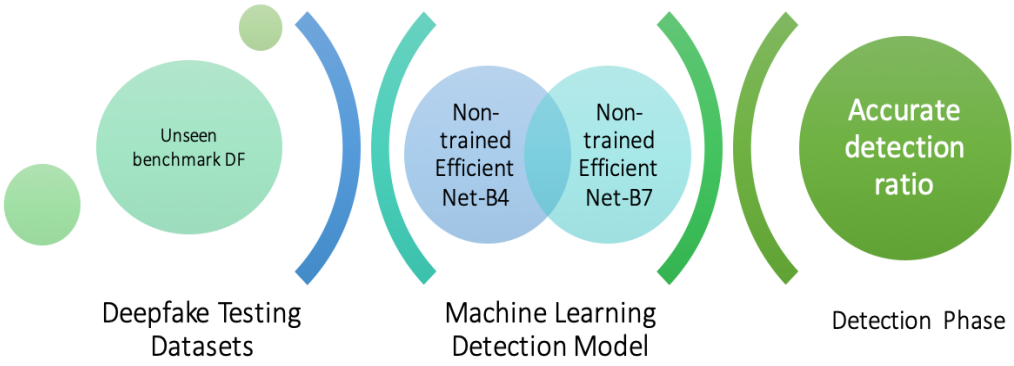
EfficientNet-B4 and EfficientNet-B7 are based upon CNNs; the base architecture of EfficientNet is CNN scaling in multiple dimensions (e.g., all dimensions of depth, width, and scale resolution together). Thus, there is an optimal balance for each dimension relative to the others (Wright, 2019). This is what makes EfficientNet more effective than other neural networks (Tan & Le, 2019).

Participants and Research Setting

In this study, no population was involved, as research is based on secondary deepfake detection datasets for both research questions. Therefore, the first phase was to answer RQ1: to locate a difference between the effectiveness of state-of-the-art CNN deepfake detection techniques, EfficientNet-B4 and EfficientNet-B7, when untrained on any datasets and validated on the same benchmark testing dataset as revealed in Figure 8. This validation test provided an accurate detection ratio of EfficientNet-B4 and EfficientNet-B7 before training.

Figure 8

Phase One of the Research



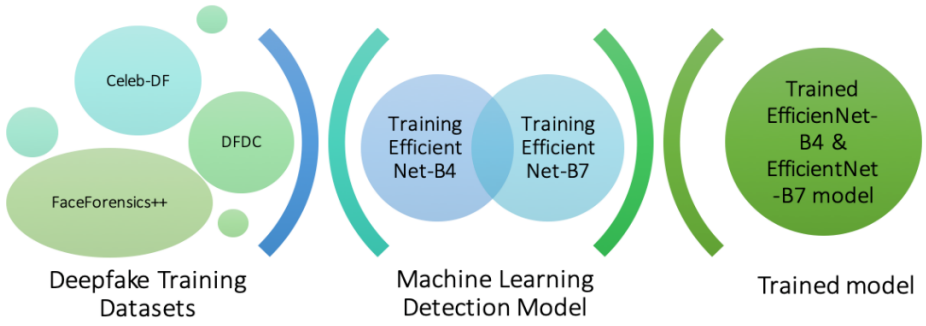
Note. The flow of the first research phase, untrained EfficientNet-B4 & EfficientNet-B7 put on unseen benchmark testing

The next phase was to answer RQ2: to locate a difference in the effectiveness of advanced CNN deepfake detection techniques, EfficientNet-B4 and EfficientNet-B7, when trained on similar datasets validated on the same benchmark testing dataset. In this phase, training was conducted for both detection technique models on Celeb-DF (Li et al., 2019), DFDC (Facebook, 2020), and FaceForensics++ (Rössler et al., 2019). Then, the validation test was conducted on EfficientNet-B4 and EfficientNet-B7 using Unseen Benchmark deepfake datasets Celeb-DF, which contains 1,692 videos (Li et al., 2019), DFDC, which includes 30,000 videos (Facebook, 2020), and the FaceForensics++ Benchmark dataset, which contains 1,000

images (Rössler et al., 2019). This validation test offered a correct detection ratio of EfficientNet-B4 and EfficientNet-B7 after training both detection models on various deepfake datasets, such as CelebDF (Li et al., 2019), DFDC (Facebook, 2020), and FaceForensics++ (Rössler et al., 2019). The process is presented in Figures 9 and 10 below.

Figure 9

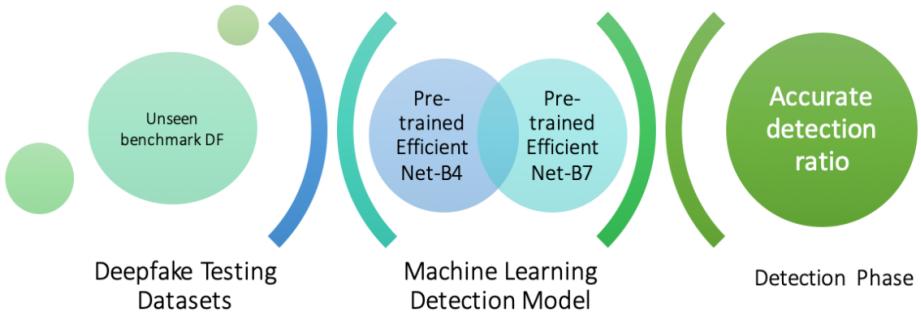
Phase Two of the Research: Training EfficientNet-B4 and EfficientNet-B7



Note. Flow of second research phase, EfficientNet-B4 & EfficientNet-B7 training on Celeb-DF, FaceForensics++ and DFDC datasets

Figure 10

Phase Two of the Research: Validating EfficientNet-B4 and EfficientNet-B7 Unseen Benchmark Dataset



Note. The flow of the second research phase, trained EfficientNet-B4 & EfficientNet-B7 put on unseen benchmark testing

For both deepfake detection techniques, the testing dataset provided an accuracy ratio of how effective these detection techniques were when seeing new deepfakes. Thus, it was a

practical approach for the comparative study of modern deepfake detection techniques to develop detection approaches further. Furthermore, a merit-based evaluation will influence the standard acceptability of deepfake detection techniques for future use.

Analyses of Research Questions

EfficientNet-B4 and EfficientNet-B7: Before Training

For phase one of the study, RQ1 discovered a difference in the effectiveness of EfficientNet-B4 and EfficientNet-B7 when untrained on any datasets. Furthermore, it validated the same benchmark testing dataset conducted by testing on the Unseen Benchmark deepfake datasets. Table 1 presents the validation test results of an accurate detection ratio of EfficientNet-B4 and EfficientNet-B7 before training them on Celeb-DF (Li et al., 2019), the DFDC (Facebook, 2020) dataset, and FaceForensics++ (Rössler et al., 2019).

Table 1

Accuracy of Detection Before Training

Model	Accuracy
EfficientNet-B4	47%
EfficientNet-B7	53%

While validating both models without training them on specific datasets and instigating the test, EfficientNet-B4's and EfficientNet-B7's achievements of accuracy in detection without training were 47% and 53%, respectively. Compared to EfficientNet-B4, EfficientNet-B7 performed slightly better; below are both models' correct and incorrect detection counts on 32,000 data points from the Unseen benchmark deepfake detection dataset.

Table 2

2 X 2 Contingency Table of EfficientNet-B4 and EfficientNet-B7 Before Training

EfficientNet-B4	EfficientNet-B7	
	Correct	Wrong
Correct	7,520	9,440
Wrong	7,520	7,520

The ratio is 9440:7520, which is less conclusive about which model is the more advantageous one to choose. The researcher followed the steps below to calculate McNemar's test for a contingency to the first research question and hypothesis, as presented in Table 2. To test the null hypothesis that the predictive performance of two models is equal (using a significance level of α=0.05), a corrected McNemar test for computing the chi-square and p-value was later conducted, as revealed in Figure 11.

Figure 11

McNemar Test for Computing the Chi-Square and P-Value

```
import numpy as np

tb_b = np.array([[7520, 9440],
                 [7520, 7520]])

from mlxtend.evaluate import mcnemar

chi2, p = mcnemar(ary=tb_b, corrected=True)
print('chi-squared:', chi2)
print('p-value:', p)


output:

chi-squared: 217.13213443396228
p-value: 3.8190599839569214e-49
Process finished with exit code 0
```

Note. RQ1 Results

Since the *p*-value surpassed our assumed significance threshold ($\alpha=0.05$), the study cannot reject the null hypothesis and assumes no significant difference in effectiveness between EfficientNet-B4 and EfficientNet-B7 when untrained on any datasets and validated on a benchmark dataset.

EfficientNet-B4 and EfficientNet-B7: After Training

For phase two of the study, RQ2 located a difference in the effectiveness of the EfficientNet-B4 and EfficientNet-B7 deepfake detection techniques trained on similar datasets and validated on the same benchmark testing dataset conducted by testing on Unseen Benchmark deepfake Datasets. Table 3 reveals the validation test results of an accurate detection ratio of EfficientNet-B4 and EfficientNet-B7 after training them on Celeb-DF (Li et al., 2019), the DFDC (Facebook, 2020) dataset, and FaceForensics++ (Rössler et al., 2019).

Table 3

2 X 2 Accuracy of Detection After Training

Model	Accuracy
EfficientNet-B4	89%
EfficientNet-B7	91%

While validating both models, training them on specific datasets, and performing the test, EfficientNet-B4's and EfficientNet-B7's achievements of the accuracy of detection without training were 89% and 93%, respectively. As demonstrated in Table 4, compared to EfficientNet-B4, EfficientNet-B7 performed slightly better; below are the correct and incorrect detection counts of both models on 32,000 data points from the Unseen benchmark deepfake detection dataset.

Table 4

2 X 2 Contingency Table of EfficientNet-B4 and EfficientNet-B7 After Training

EfficientNet-B4	EfficientNet-B7	
	Correct	Wrong
Correct	26,699	601
Wrong	1	2,699

The ratio is 601:1, which is less conclusive about which model is the better one to choose. To test the second research question and hypothesis, the steps to calculate McNemar's test for contingency are revealed in Table 4. To test the null hypothesis that the predictive performance of two models is equal (using a significance level of $\alpha=0.05$), a corrected McNemar Test for computing the chi-square and *p*-value is later conducted, as revealed in Figure 12.

Figure 12

McNemar Test for Computing the Chi-Square and P-Value

```
import numpy as np

tb_b = np.array([[26699, 601],
                 [1, 2699]])

from mlxtend.evaluate import mcnemar

chi2, p = mcnemar(ary=tb_b, corrected=True)
print('chi-squared:', chi2)
print('p-value:', p)
output

chi-squared: 596.014950166113
p-value: 1.2319238348649218e-131

Process finished with exit code 0
```

Note. RQ2 Results

Since the *p*-value surpassed the assumed significance threshold ($\alpha=0.05$), the researcher could not reject the null hypothesis and assume no significant difference in effectiveness

between the EfficientNet-B4 and EfficientNet-B7 techniques when trained on any datasets validated on a benchmark dataset.

Summary

Chapter 4 has focused on presenting comparative study findings in detection techniques and their effectiveness based on unseen training datasets to measure correct detection ratios. The Researcher tested EfficientNet-B4 and EfficientNet-B7 deepfake detection models on the Unseen benchmark dataset for two different scenarios: one with no training models and another for training both models of the latest deepfake datasets. As a result of the experiment, the null hypotheses in both research questions were rejected, as there was no significant difference in performance. In the final chapter, a deep dive into this dissertation's research questions is presented alongside a practical assessment regarding them.

Chapter Five: Summary, Discussion, and Implications

Introduction

The purpose of this study was to perform a comparative analysis of the EfficientNet-B4, and EfficientNet-B7 neural network-based AI deepfake detection techniques and determine their performance effectiveness. An experiment was conducted with EfficientNet-B4 and EfficientNet-B7 models to detect deepfakes in the Unseen benchmark deepfake dataset, testing both models with two different scenarios. In the first scenario, EfficientNet-B4 and EfficientNet-B7 were not trained on any datasets. In the second scenario, EfficientNet-B4 and EfficientNet-B7 were trained on three deepfake datasets.

The first scenario's purpose was to replicate a real-time situation, where many new deepfake detection algorithms and techniques are being developed. What if we were to plug those algorithms directly onto live platforms to detect the latest trends of deepfake on the go?

In the second phase of the study, EfficientNet-B4 and EfficientNet-B7 models were trained on specific datasets, considering where active-active development is happening for deepfake detection techniques and where datasets can mimic and perform supervised learning those machine learning models. This study has aimed to help determine the current and future capabilities of deepfake technologies and their scope and threat to national security.

Practical Assessment of Research Questions

Deepfake detection development prioritizes big-tech companies and defense agencies, as deepfake has become a threat to national security and democracy. This quantitative study has aimed to evaluate the effectiveness of different deepfake detection techniques based on neural networks. Findings drive efforts in creating, developing, and brainstorming detection techniques; it becomes more crucial to validate the rapid development of detection techniques to improve

and increase other deepfake detection accuracy ratios. In this study, EfficientNet-B4 and EfficientNet-B7, two advanced deepfake detection machine learning models, were chosen.

To validate the performance of these deepfake detection models, this study has used an unseen benchmark dataset. It was created as part of this study by combining three different datasets, that is, Celeb-DF (Li et al., 2019), DFDC (Facebook, 2020), and FaceForensics++ (Rössler et al., 2019) by reserving a set of 30% videos from Celeb-DF (Li et al., 2019) and the DFDC (Facebook, 2020) dataset alongside the FaceForensics++ (Rössler et al., 2019) benchmark dataset of 1,000 images.

Machine learning models learn by neural network layers and training epochs. Therefore, this study was performed in two experiment phases to answer both research questions:

For RQ1, There was no significant difference between the effectiveness of EfficientNet-B4 and EfficientNet-B7 when not trained on datasets and validated on the benchmark dataset. The deepfake models did not perform well on unseen deepfake datasets, concluding that if deepfake detection models were nontrained, they might be unreliable to detect deepfakes in real time.

For RQ2, The deepfake models were pre-trained on Celeb-DF, FaceForensics++, and DFDC datasets and validated on the benchmark dataset. Both models performed well because they were pre-trained. But there was no significant difference between the effectiveness of EfficientNet-B4 and EfficientNet-B7. If deepfake detection models had exposure to deepfake datasets, performance improved on the same scale, making both deepfake models reliable. Not one model is better than the other. The accurate detection ratio was the same for both models.

The first scenario's purpose was to replicate a real-time situation, where many new deepfake detection algorithms and techniques are being developed. What if we were to plug

those algorithms directly into live platforms to detect the latest trends of deepfake on the go? In the second phase of the study, EfficientNet-B4 and EfficientNet-B7 models were trained on specific datasets, considering where active-active development is happening for the deepfake detection technique, and datasets are available to train and perform supervised learning for those machine learning models can be done.

Result of RQ1: When EfficientNet-B4 and EfficientNet-B7 were tested directly on the Unseen benchmark datasets, both models performed poorly. The poor performance was that the models did not detect any differences between authentic and fake datasets. Both the EfficientNet-B4 and EfficientNet-B7 models had never encountered any deepfake content. Algorithms require conditional flow and decision-making ability for any input to the models. In this case, exceptions occurred, and the neural networks failed to detect and create real/fake labels for the dataset images.

Result of RQ2: When EfficientNet-B4 and EfficientNet-B7 trained on Celeb-DF (Li et al., 2019), DFDC (Facebook, 2020), and FaceForensics++ (Rössler et al., 2019) with specified epochs, both models performed to relatively more outstanding abilities. Models were fit for detecting differences between real and fake images from unseen datasets for most of the data points for some performance improvement. The EfficientNet-B4 and EfficientNet-B7 models encountered deepfake and actual content during training; both models presented a learning direction. They can identify or label discovered content as real or fake, making accurate predictions on unseen datasets. In this case, exceptions occurred where EfficientNet-B4 and EfficientNet-B7 models could not identify deepfake or actual images when off-trend content was encountered.

Limitations of the Study

EfficientNets are based on CNNs architecture and scaling methods that uniformly scale all depth, width, and resolution dimensions using a compound co-efficient. In this study, EfficientNet-B4 and EfficientNet-B7 trained on epochs set to 100. If you train a model on smaller datasets for more epochs, then machine learning models can be overfitted, which limits these models. Over fitted, EfficientNet-B4, and EfficientNet-B7 use more noise, increasing their performance in cases of known noise (training data) and decreasing their performance in novel noise cases (testing data).

In this study, overfitting a model is an actual limitation, as datasets used to train the EfficientNet-B4 and EfficientNet-B7 were highly focused on images and videos of people; in each image, a person was present, whereas deepfake concentrates on facial features.

Currently, detecting deepfake videos is often formulated, solved, and evaluated as a binary classification problem, where each video is categorized as real or deepfake. However, such dichotomy is simple to align in controlled experiments, where we develop and test deepfake detection algorithms using videos that are either pristine or made with deepfake generation algorithms.

However, the image is murkier when the detection methods are deployed on the planet. For example, videos may be fabricated or manipulated in ways different from deepfakes. Therefore, not being detected as a deepfake video does not necessarily suggest that the video is real. Additionally, a deepfake video may be subject to other kinds of manipulations, and one label might not comprehensively reflect this.

Furthermore, during a video with multiple subjects' faces, just one or some are generated with deepfake for a fraction of the frames so that the binary classification scheme must be

extended to multiclass, multi-label, and native classification/detection to handle the complexities of planet media imitations (Lyu, 2020). As a result, EfficientNet-B4 may outperform EfficientNet-B7 on dataset one Celeb-DF (Li et al., 2019) and dataset two the DFDC (Facebook, 2020) separately. However, EfficientNet-B7 can still exceed EfficientNet-B4 overall. This all depends on how the models are trained or when the evaluation matrix was taken. This is also a sign of the models' overfitting.

Implications for Future Study

The result of this study has been a learning curve. We must regularly evaluate EfficientNet machine learning models as deepfake creation trends shift to new trends and techniques become faster than detection. It would be a crucial breakthrough if the evaluation of state-of-the-art deepfake detection techniques, EfficientNet-B4 and EfficientNet-B7, could be achieved in an active-active manner, where models can learn and train more quickly on novel noises.

EfficientNet-B4 and EfficientNet-B7 are both powerful techniques to detect deepfake content. This study's findings can help improve the effectiveness of EfficientNet-B4 and EfficientNet-B7. This can be enhanced if both methods work together as a mixed-approach detection, which can be more reliable and accurate, as there is no significant difference in the performance of both detection techniques. Critical factors for better detection techniques are additional parameters and labels, not just real or fake content classification. In the future, creating deepfake detection based on features can help provide authenticity of the content and filter out the deepfake content before it reaches users.

Summary

Chapter 5 has presented the research findings regarding the EfficientNet-B4 and EfficientNet-B7 deepfake detection techniques. On the practical assessment of research, if they are trained correctly with advantageous dataset quality, CNNs model can be combined to outperform individual detection capabilities of EfficientNet-B4 and EfficientNet-B7. Furthermore, no significant differences were determined in the effectiveness of these techniques, which exists as a breakthrough. Therefore, both EfficientNet-B4 and EfficientNet-B7 can be reliable deepfake detection techniques for a given time.

References

Afchar, D., Echizen, I., Nozick, V., and Yamagishi, J. (2018) "Mesonet: a compact facial video forgery detection network," in IEEE International Workshop on Information Forensics and Security (WIFS).

Agarwal, S., Farid, H., Gu, Y., He, M., Li, H., and Nagano, K. "Protecting world leaders against deep fakes," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 38–45

Agarwal, Vardan. "Complete Architectural Details of All EfficientNet Models." *Medium*, 31 May 2020, towardsdatascience.com/complete-architectural-details-of-all-efficientnet-models-5fd5b736142. Accessed 29 Nov. 2020.

Andrey, Sam. "deepfake: Everything You Need to Know About It." *Medium*, 11 Dec. 2020, medium.com/swlh/deepfake-everything-you-need-to-know-about-it-3aa8a0ab6c05. Accessed 16 Apr. 2021.

Anderson, Janna, and Lee Rainie. "Concerns about Democracy in the Digital Age." *Pew Research Center: Internet, Science & Tech*, 21 Feb. 2020, www.pewresearch.org/internet/2020/02/21/concerns-about-democracy-in-the-digital-age/.

Chung, J., Senior, A., Vinyals, O., & Zisserman, A. (2016) "Lip reading sentences in the wild," CoRR, vol. abs/1611.05358 [Online]. Available: http://arxiv.org/abs/1611.05358

"deepfake Video: It Takes AI to Beat AI." *HPE*, 2019, www.hpe.com/us/en/insights/articles/deepfake-video-it-takes-ai-to-beat-ai-1911.html

Facebook. (2020, March 25). Deepfake Detection Challenge Dataset. Retrieved July 12, 2020, from ai.facebook.com website: https://ai.facebook.com/datasets/dfdc/

Farid, H. (2012) Digital Image Forensics, MIT Press.

Güera, D., & Delp, E. (2018). D*eepfake Video Detection Using Recurrent Neural Networks*. Retrieved from https://engineering.purdue.edu/~dgueraco/content/deepfake.pdf

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. Generative Adversarial Networks. In NIPS, 2014.

Griffin, M. (2019, August 2). Edtech company Udacity uses deepfake tech to create educational videos automatically. Retrieved July 11, 2020, from By Matthew Griffin Futurist and Keynote Speaker website: https://www.fanaticalfuturist.com/2019/08/edtech-company-udacity-uses-deepfake-tech-to-create-educational-videos-automatically/

IBM Corp. Released 2020. IBM SPSS Statistics for Macintosh, Version 27.0. Armonk, NY: IBM Corp. Retrieved from https://hadoop.apache.org

Iizuka, S., Simo-Serra, E., & Ishikawa, H. Globally and locally consistent image completion. ACM Trans. Graph., 36(4):107:1–107:14, 2017.

Jaiman, A. (2020, July 7). Deepfake detection is super hard! Retrieved from Medium website: https://towardsdatascience.com/deepfake-detection-is-super-hard-38f98241ee49

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017) "Progressive growing of gans for improved quality, stability, and variation," The International Conference on Learning Representations (ICLR).

Karras, T., Laine, S., and Aila, T. "A style-based generator architecture for generative adversarial networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.

Khodabakhsh, A., Ramachandra, R., Raja, K., Wasnik, P., and Busch, C. "Fake face detection methods: Can they be generalized?" in 2018 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, 2018, pp. 1–6.

Kingma, D. P., & Welling, M. Auto-encoding variation Bayes. In ICLR, 2014.

Knight, W. (2019, August 6). Facebook, Google, Twitter are not prepared for presidential deepfakes. Retrieved from MIT Technology Review website: https://www.technologyreview.com/f/614087/facebook-google-twitter-arent-prepared-for-presidential-deepfakes/

Korshunov, P., & Marcel, S. (2018) *DeepFakes: A New Threat to Face Recognition? Assessment and Detection*.

Korshunov, P., and Marcel, S. "Speaker inconsistency detection in tampered video," in 2018 26th European Signal Processing Conference (EUSIPCO). IEEE, 2018, pp. 2375–2379.

Li, Y., Chang, M., and Lyu, S. (2018) "Exposing AI-generated fake face videos by detecting eye blinking," in IEEE International Workshop on Information Forensics and Security (WIFS).

Li, Y., & Lyu, S. (2019). *Exposing DeepFake Videos by Detecting Face Warping Artifacts*. Retrieved from https://openaccess.thecvf.com/content_CVPRW_2019/papers/Media%20Forensics/Li_Exposing_DeepFake_Videos_By_Detecting_Face_Warping_Artifacts_CVPRW_2019_paper.pdf

Liu, M., Breuel, T., & Kautz, J. (2017) Unsupervised image-to-image translation networks. CoRR, abs/1703.00848.

Liu, Z., Song, G., Cai, J., Cham, T-J., and Zhang, J. "Conditional adversarial synthesis of 3d facial action units," Neurocomputing, vol. 355, pp. 200–208, 2019.

Lyu, S. (2020) *deepfake detection: current challenges and next steps*.

Majumdar, P., Agarwal, A., Singh, R., and Vatsa, M. "Evading face recognition via partial tampering of faces," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0

Martin, Gary. 'A Picture Is Worth a Thousand Words' - the Meaning and Origin of This Phrase." *Phrasefinder*, 2019, www.phrases.org.uk/meanings/a-picture-is-worth-a-thousand-words.html.

Melas-Kyriazi, Luke. "Lukemelas/EfficientNet-PyTorch." *GitHub*, 30 Oct. 2019, github.com/lukemelas/EfficientNet-PyTorch.

Nguyen, T., Nguyen, C., Nguyen, D., Nguyen, T., & Nahavandi, S. (2020). D*eep Learning for deepfakes Creation and Detection*. Retrieved from https://arxiv.org/pdf/1909.11573.pdf

Rössler, A, et al. *FaceForensics++: Learning to Detect Manipulated Facial Images*. 2019.

Salimans, T., Goodfellow, J., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016) Improved training GANs. In NIPS.

Seferbekov, S. (2020, June 27). selimsef/dfdc_deepfake_challenge. Retrieved June 27, 2020, from GitHub website: https://github.com/selimsef/dfdc_deepfake_challenge

Shanoob. "Deep Fake – the Greatest Threat to the Idea of Truth." *IAS EXPRESS*, 24 Nov. 2020, www.iasexpress.net/deep-fake-the-greatest-threat-to-the-idea-of-truth/. Accessed 16 Apr. 2021.

Silbey, J., & Hartzog, W. (2019). The Upside of Deep Fakes. *Maryland Law Review*, *78*(4). Retrieved from

https://digitalcommons.law.umaryland.edu/cgi/viewcontent.cgi?article=3837&context=mlr

"Singapore Introduces New Bill to Punish Online Fake News Outlets." *Coda Story*, 3 Apr. 2019, www.codastory.com/news/singapore-introduces-new-bill-punish-online-fake-news-outlets/. Accessed 16 Apr. 2021.

Singh, G. "deepfake and AI: To Be or Not to Be." *Copperpod IP*, 11 Feb. 2020, www.copperpodip.com/post/2020/02/11/deepfake-and-ai-to-be-or-not-to-be.

Suciu, P. (2020, January 21). Know Your Enemy: The Difficulty of Defining Deepfakes. Retrieved April 21, 2021, from www.technewsworld.com website: https://www.ecommercetimes.com/story/86466.html

Suwajanakorn, S., Seitz, M., & Kemelmacher-Shlizerman, I. (2017) "Synthesizing Obama: Learning lip sync from audio," ACM Trans. Graph., vol. 36, no. 4, pp. 95:1–95:13.

Tan, M., & Le, Q. (2019a). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. Retrieved from https://arxiv.org/pdf/1905.11946.pdf

Tan, M., & Le, Q. V. (2019b, May 29). EfficientNet: Improving Accuracy and Efficiency through AutoML and Model Scaling. Retrieved August 10, 2020, from Google AI Blog website: https://ai.googleblog.com/2019/05/efficientnet-improving-accuracy-and.html

Van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A. (2016a) ¨ Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. CoRR, abs/1609.03499.

Van den Oord, A., Kalchbrenner, N., & Kavukcuoglu, K. (2016b) Pixel recurrent neural networks. ¨ In ICML, pp. 1747–1756.

Van den Oord, A., Kalchbrenner, N., Vinyals, O., Espeholt, L., Graves, A., & Kavukcuoglu, K. (2016c) Conditional image generation with PixelCNN decoders. CoRR, abs/1606.05328.

Wagner, C., & Henderson, K. (2020). *Technology: "Deepfakes."* Retrieved from website: https://helenacollege.edu/library/docs/wagner_essay2_technology.pdf

Wang, T., Liu, M., Zhu, J., Tao, A., Kautz, J., & Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional GANs. CoRR, abs/1711.11585, 2017.

Wang, Y., Dantcheva, A. A video is worth more than 1000 lies. Comparing 3DCNN approaches for detecting deepfakes. FG'20, 15th IEEE International Conference on Automatic Face and Gesture Recognition, May 18-22, 2020, Buenos Aires, Argentina., May 2020, Buenos Aires, Argentina. ffhal-02862476f

Westerlund, M. (2019). The Emergence of deepfake Technology: A Review. *Technology Innovation Management Review*, *9*(11). Retrieved from https://timreview.ca/article/1282

Wright, L. (2019, June 1). EfficientNet from Google — Optimally Scaling CNN model architectures with "compound scaling." Retrieved April 21, 2021, from Medium website: https://lessw.medium.com/efficientnet-from-google-optimally-scaling-cnn-model-architectures-with-compound-scaling-e094d84d19d4

Wubet, W. M. (2020). The deepfake Challenges and deepfake Video Detection. *International Journal of Innovative Technology and Exploring Engineering*, *9*(6), 789–796. https://doi.org/10.35940/ijitee.e2779.049620

Yang, X., Li, Y., & Lyu, S. (2019) "Exposing deep fakes using inconsistent head poses," in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Zhu, J., Park, T., Isola, P., & Efros, A. A. (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. CoRR, abs/1703.10593.